band spectrum in the treatment of the behavior of an electron in a crystal under the influence of an external force. The following is the resulting detailed picture:

Let us assume that the electron is, for instance, in a quantum state in the middle of a band. Its eigenfunction is then essentially a *running* wave, and the electron transports a charge which constitutes a current. If a force acts in the direction of propagation of the wave, the energy of the electron is increased and it must change into a higher quantum state. After a while it occupies the highest quantum state in the particular band, namely, the upper band edge. In this state, the wavelength corresponds to the Bragg reflection condition for a certain group of lattice planes. The in-phase superposition of the spherical waves diffracted by the lattice points results in a wave propagated in the opposite direction with the same intensity, so that the electron in this quantum state is represented by a *standing* wave. The electron does not transport any charge in this state and, therefore, does not constitute a current. In spite of the energy increase, the electron has, as a result of the participation of the lattice, lost its transport action and its velocity. The electron has been slowed down to zero velocity, although the force and propagation direction of the wave, i.e., the force and electron velocity, were in the same direction.

The behavior of an electron in a crystal under the influence of the lattice is compared with a "quasi-free" electron, which is not affected by lattice forces, by describing the process on one hand by the equation

$$\mathbf{F} + \text{lattice forces} = m \cdot \frac{d\mathbf{v}}{dt} \qquad \text{(III.2.01)}$$

and on the other hand by the equation

$$\mathbf{F} = m_{\text{eff}} \cdot \frac{d\mathbf{v}}{dt} \qquad \text{(III.2.02)}$$

According to (III.2.02) one obtains the correct value for the acceleration $d\mathbf{v}/dt$ only if the effective mass m_{eff} of the quasi-free electron has a value[1] m_n which is different from the real mass m. In the case just considered, m_n must even be negative.

In general we find the following value[2] for m_n

$$m_n = \frac{\hbar^2}{E_n''(k, N)} \qquad \text{(III.2.03)}$$

[1] To distinguish it from the effective mass m_p of holes to be introduced later.

[2] See in this connection Chap. VII, §5 and §6, particularly Eq. (VII.6.28), and D. Pfirsch and E. Spenke, *Z. Physik*, **137**: 309–312 (1954).

where E_n is the energy of the *n*egative electron. E_n depends on the quantum state of the electron and is, therefore, a function of the wave number $k = 2\pi/\lambda$ of the electron (λ = de Broglie wavelength). In view of the band structure, E_n is a multivalued function of k (see Fig. III.2.1). The various branches, or more specifically "bands," are characterized by the band number N. "′" indicates differentiation with respect to the wave number k.

It is now evident that in the upper part of the band the electron behaves like a free electron with a negative mass. Experience has shown that this statement is at first somewhat difficult to comprehend.

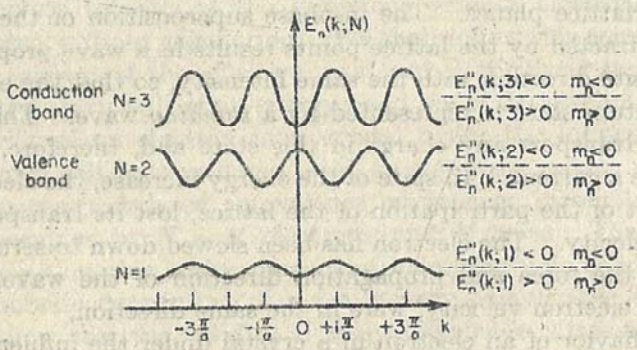

FIG. III.2.1. Electron energy E_n as a function of the wave number k. In the lower parts of the bands m_n is positive and in the upper parts it is negative.

$$[E_n''(k;N) = \frac{\partial^2}{\partial k^2} E_n(k;N)]$$

I hope that I have helped to overcome this obstacle by pointing out as clearly as possible that the definition of the "effective" mass considers only the external force **F** and neglects entirely the strong influence of the lattice on the electron. It is not surprising that, with this approach, the proportionality factor between force and acceleration assumes most peculiar, e.g., negative, values.

However, for many considerations it is convenient to treat the totality of the electrons as a *free* electron gas. If this can be accomplished by a simple substitution of an effective mass for the real value of the electron mass, it is worth while to accept the modified mass values which can even become negative in the upper part of a band. In order to remember that the electrons are not really free but are subject to exceedingly strong lattice influences which are only apparently eliminated by an artifice, one often speaks of *quasi-free* electrons.

§3. The Equivalence of an Almost Filled Valence-Band with a Fermi Gas of Quasi-free Holes

We shall now undertake the proof of the equivalence between an incompletely filled valence band and a Fermi gas of quasi-free holes. A logical step is the comparison of the following two particle ensembles:

1. $N - M$ electrons which leave M states of the N possible valence-band states unoccupied. We shall call this quantity of particles the "real almost filled valence band."

2. M fictitious positive charges $(+e)$ together with N ordinary negative electrons $(-e)$ which just fill the N available states of a valence band.[1] We shall call this quantity of particles the "filled valence band with fictitious positive additional charges."

First we shall consider the action of an external electric field on the two particle ensembles to be compared. The current thus produced is the same in both cases, provided that we can assign such properties to the M *fictitious* charges $+e$ that they move under the influence of the external electric field (and the lattice forces) in the same manner as those M electrons of the filled band which are in the M quantum states that are left unoccupied in the real valence band. This identical course in space and time of the M electrons and M fictitious positive charges can be attained simply by assigning a mass $-m$ to the positive charges. The external field as well as the lattice potential exert exactly opposite forces on the positive additional charges as on the electrons since the signs of the charges are opposite. Exactly the same motion results, however, if the sign of the mass is changed too.

From the standpoint of wave mechanics, the identical motion of the accompanied electron $(-e, +m)$ and the accompanying additional positive charge $(+e, -m)$ may be established as follows. The probability amplitude ψ_n of the *negative* electron satisfies the Schrödinger equation

$$-\frac{\hbar^2}{2m}\Delta\psi_n - eU(\mathbf{r})\psi_n = +j\hbar\frac{\partial\psi_n}{\partial t}; \qquad U(\mathbf{r}) = \text{lattice potential} \tag{III.3.01}$$

The probability amplitude ψ_p of the fictitious *positive* additional charge satisfies the Schrödinger equation

[1] In this book e is always the absolute value $+1.6 \cdot 10^{-19}$ coulomb of the elementary charge, so that the charge of the negative electron is $-e$ and that of the positive hole $+e$.

$$+\frac{\hbar^2}{2m}\Delta\psi_p + eU(\mathbf{r})\psi_p = +j\hbar\frac{\partial\psi_p}{\partial t} \qquad \text{(III.3.02)}$$

Comparison shows that the probability amplitudes ψ_n and ψ_p are not directly equal, but that they are complex conjugates.

$$\psi_n(x;\, k,\, N) = \psi_p^*(x;\, k,\, N) \qquad \text{(III.3.03)}$$

k = wave number; N = band number.

This is sufficient for the exact agreement of the space-time distribution of the probabilities $\psi_n\psi_n^*$ and $\psi_p\psi_p^*$, respectively, so that the positive charge accompanies the electron continually. If one converts to the stationary Schrödinger equations

$$-\frac{\hbar^2}{2m}\Delta\psi_n - eU(\mathbf{r})\psi_n = E_n\psi_n \qquad \text{(III.3.04)}$$

$$+\frac{\hbar^2}{2m}\Delta\psi_p + eU(\mathbf{r})\psi_p = E_p\psi_p \qquad \text{(III.3.05)}$$

or

$$+\frac{\hbar^2}{2m}\Delta\psi_p^* + eU(\mathbf{r})\psi_p^* = E_p\psi_p^* \qquad \text{(III.3.06)}$$

it can be seen that

$$E_n(k,\, N) = -E_p(k,\, N) \qquad \text{(III.3.07)}$$

This relation will be useful later.

The field thus creates the same current, on one hand, in the real almost filled valence band and, on the other hand, in the filled valence band with the fictitious positive charges of mass $-m$. If we calculate the current for the latter particles alone, the contribution of the filled valence band is eliminated; for we have already pointed out in the introduction that in a filled valence band the current contribution of the electrons in the lower part of the band with the positive effective masses and the current contribution of the upper part of the band with the negative effective masses compensate each other exactly. Therefore, the only remaining current is the original contribution of the M positive charges, namely, the "holes" with the charge $+e$ and the mass $-m$.

Similar to the case of the electrons in §2, we can now change over from the holes which are subject to the lattice forces (holes in a crystal) to the *quasi-free* holes by introducing an effective mass value m_p in place of $-m$, thus taking into account the effect of the lattice forces. In analogy to Eq. (III.2.03), we find

$$m_p = \frac{\hbar^2}{E_p''(k;\, N)} \qquad \text{(III.3.08)}$$

and with the help of (III.3.07) we obtain

$$m_p(k;\, N) = -m_n(k;\, N) \qquad \text{(III.3.09)}$$

The properties of electrons and holes are summarized once more in Fig. III.3.1.

The distribution of the M holes among the quantum states of the valence band is governed in the foregoing considerations by the following rule: Each quantum state which is not occupied in the real almost filled valence band can accommodate one hole in the filled valence band with a fictitious positive additional charge. This rule can now be

	Electrons	Holes
Charge	$-e$	$+e$
Mass	$+m$	$-m$
Effective mass at the *upper* edge of the *valence band*	$m_n = \dfrac{\hbar^2}{E''(k,\text{ valence})} < 0$	$m_p = -\dfrac{\hbar^2}{E''(k,\text{ valence})} > 0$
Effective mass at the *lower* edge of the *conduction band*	$m_n = \dfrac{\hbar^2}{E''(k,\text{ conduction})} > 0$	$m_p = -\dfrac{\hbar^2}{E''(k,\text{ conduction})} < 0$

FIG. III.3.1. Properties of electrons and holes.
$E(k,$ number of band$) = E_n(k, \ldots) =$ energy of the *electrons*.

replaced by one in which only holes are considered without reference to the distribution of electrons. This comes about in the following manner:

The probability that a quantum state with the electron energy E_n is occupied by an electron is given by the Fermi distribution function

$$f(E_n) = \frac{1}{e^{\frac{E_n - E_F}{kT}} + 1}$$

The probability that this state is *not* occupied with an electron is therefore

$$1 - f(E_n) = \frac{e^{\frac{E_n - E_F}{kT}} + 1 - 1}{e^{\frac{E_n - E_F}{kT}} + 1} = \frac{e^{\frac{E_n - E_F}{kT}}}{e^{\frac{E_n - E_F}{kT}} + 1} = \frac{1}{1 + e^{-\frac{E_n - E_F}{kT}}}$$

$$= \frac{1}{e^{\frac{(-E_n) - (-E_F)}{kT}} + 1} = \frac{1}{e^{\frac{E_p - (-E_F)}{kT}} + 1} = f(E_p)$$

At the end we have put $-E_n = E_p$. Besides, the energy of the Fermi level changes sign *as well*, like all energy values. E_p is then the energy of the positive hole; for in the potential or electrostatic part of the energy we find the charge as a factor, and in the kinetic part the mass. Both factors change their sign when we change over from electron to hole.[1]

The original rule yields the probability $1 - f(E_n)$ for the occupancy of a quantum state by a hole. Without reference to electrons, we find the same probability of occupancy if we postulate Fermi statistics for the holes because then we have to introduce $f(E_p)$ as the probability of occupancy of a quantum state. This is equal to $1 - f(E_n)$, according to the foregoing derivation.

In summary we see that the current contribution of the real valence band which is almost filled with $(N - M)$ electrons has exactly the same magnitude as that of a Fermi gas of M quasi-free holes with the charge $+e$ and the effective mass $m_p = -\hbar^2/E''(k, \text{valence}) > 0$.

The preceding argument can certainly be extended to the transport phenomena where magnetic fields $\mathbf{H}$ act upon the electrons. The Lorentz force $-(e/c)\mathbf{v} \times \mathbf{H}$ which acts upon an electron moving with the velocity $\mathbf{v}$ contains also the charge as proportionality factor. Hence here, too, we must assign the mass $-m$ to the fictitious positive additional charges in order to compensate for the change of sign of the charge; the subsequent conclusions are the same as in the case of the electric fields.

Transport phenomena are not necessarily always a consequence of electric or magnetic fields; they can also be caused by concentration or temperature gradients. The pertinent relations follow from the statistics of the participating current carriers. We have just shown that the statistics of negative electrons lead to the same results as the Fermi statistics of positive holes, so that we are justified in applying the hole representation also to transport phenomena which are caused by concentration and temperature gradients.

Finally, if the conductor is accelerated or decelerated, the electrons in a conductor are set in motion by forces of inertia, similarly to soup which spills over in a carelessly moved plate.

If a coordinate system is, for instance, accelerated in the positive x direction, a mass M which is not acted upon by any forces will remain unmoved and stay behind with respect to the coordinate system. To the observer, moving with the acceleration, the mass M appears to encounter a force in the direction of the negative x axis. If the mass

[1] $-E_n = E_p$ also followed from the comparison of the two Schrödinger equations for electrons and holes. See Eq. (III.3.07).

M encounters other additional forces (such as springs under tension) and the mass M performs certain motions due to these spring forces in the reference system, the observer in the accelerated reference system sees a change of these motions as if the mass M encountered an inertial force in the direction of the negative x axis in addition to the spring forces.

We shall apply these generally known relations to the accelerated solid and its electrons in the Tolman experiment: If a conductor encounters an acceleration $+\mathbf{g}$, a force $\mathbf{F} = -m\mathbf{g}$ seems to act on its electrons. This inertial force can be replaced by an electric field $\mathbf{E} = +(m/e)\mathbf{g}$; for this field would exert a force $\mathbf{F} = -e\mathbf{E} = -m\mathbf{g}$ on the electrons with their negative charges $-e$. The equivalent field $\mathbf{E}$ would, however, cause a current $\mathbf{i} = \sigma\mathbf{E}$. Therefore the acceleration $+\mathbf{g}$ of the conductor must cause a current arising from the inertial force $-m\mathbf{g}$ with a current density $\mathbf{i} = +(m/e)\sigma\mathbf{g}$, as was observed in experiments carried out by Tolman and others.

Even for such transport phenomena effected by inertial forces, we can choose either the electron or the hole representation. We have just seen that the acceleration or deceleration of a conductor acts on the observer, who is carried along, as an additional gravitational field $-\mathbf{g}$. In the same manner as one calculates the force exerted on an electron by an electric field $\mathbf{E}$, namely, by multiplication with the charge $(-e)$ one obtains the inertial force of the gravitational field $-\mathbf{g}$, namely, by multiplying with the real mass $(+m)$; for the lattice forces do not enter the picture in the calculation of the inertial force. Hence there is no possibility of overlooking them, and m_{eff} does not play any part. Accordingly, one derives the inertial force on the holes from the gravitational field $-\mathbf{g}$ by multiplication with the "true" mass $-m$. The gravitational field exerts, therefore, the inertial force $(+m)\cdot(-\mathbf{g}) = -m\mathbf{g}$ on the electrons and $(-m)\cdot(-\mathbf{g}) = +m\mathbf{g}$ on the holes.

If we calculate now the acceleration from the force, we must consider the lattice forces; this is accomplished by using the effective masses as proportionality factors between force and acceleration instead of the true masses (page 10). These have opposite signs for an electron and a hole in the same quantum state (see Fig. III.3.1). The resulting acceleration is again the same for both particles because the inertial force as well as the effective mass change their sign. This assures the permanent identity of the location in space of electron and hole. The hole representation must therefore be also applicable for the action of gravitational and inertial forces.

We have shown that the electron as well as the hole representation

is applicable to transport phenomena which are caused by electric or magnetic fields, by temperature or concentration gradients, and by inertial or gravitational forces. In any specific case it is desirable to choose the representation which yields the lower carrier number because this will tend to allow the use of the Maxwell-Boltzmann limiting case of the Fermi-Dirac statistics.

For the case of a sparsely occupied conduction band, one prefers the electron representation which leads to a Maxwell-Boltzmann gas of quasi-free negative electrons with positive effective mass $m_{e\,\mathrm{eff}}$, for $m_{\ominus\,\mathrm{eff}} = \hbar^2/E''$ (k, lower edge of conduction band) $= m_n > 0$ is positive in the lower part of the conduction band.

In the case of an almost filled valence band, however, one will choose the hole representation with which one obtains a Maxwell-Boltzmann gas of quasi-free positive holes with positive effective mass

$$m_{\oplus\,\mathrm{eff}} = -\frac{\hbar^2}{E''(k,\ \text{upper edge of valence band})} = m_p > 0$$

for E'' (k, valence) is negative at the upper edge of the valence band (see Fig. III.2.1).

Before concluding this §3, we shall briefly discuss another question. Sometimes the argument arises as to whether it is possible to distinguish between electron and hole conduction in a certain experiment. We can point out a relatively simple method to answer this question. The final equation for the result of the experiment in question should be, as far as possible, expressed in terms[1] of e, m, m_n, and m_p. Then we pass over from the hole conduction case p to the electron conduction case n by the following substitution[2]

$$\begin{aligned} p &\rightarrow n \\ +e &\rightarrow -e \\ -m &\rightarrow +m \\ +m_p &\rightarrow +m_n \end{aligned}$$

[1] Here we often utilize the equation $\mu = (e/m_{e\,\mathrm{eff}})\tau$ for the carrier mobility μ. See Eq. (VII.9.25).

[2] The validity of the first three substitutions is beyond doubt. With respect to the fourth substitution, we must realize that we are *not* dealing with optional descriptions of one and the same case—such as electrons at the lower edge of the conduction band—in the electron or hole language. The following *equation* would be applicable if this were so:

$$m_{\oplus\,\mathrm{eff}} = -m_{\ominus\,\mathrm{eff}}$$

It is rather a transition between two different cases, namely, the transition from holes at the upper edge of the valence band to electrons at the lower edge of the conduction band. The correct *substitution* for this is

$$m_{\oplus\,\mathrm{eff}} = m_p \rightarrow m_{\ominus\,\mathrm{eff}} = m_n$$

In this transition $p \rightarrow n$, the result of the experiment either changes its sign or it does not. Accordingly, the experiment either allows a differentiation between hole and electron conduction or it does not. A few examples will clarify this procedure.

1. *Electric current due to an electric field.*

Experimental result: $\mathbf{i} = \sigma \mathbf{E}$

Expressed in terms of e, m, m_n, m_p: $\mathbf{i} = e\mu_p p\mathbf{E} = \dfrac{e^2}{m_p} p\tau\mathbf{E}$

Transition $p \rightarrow n$: $\mathbf{i} = \dfrac{(-e)^2}{(+m_n)} n\tau\mathbf{E} = \dfrac{e^2}{m_n} n\tau\mathbf{E}$

There is no change of sign; therefore, it is not possible to distinguish between p-type and n-type.

2. *Hall effect.*

Experimental result: $\Theta_p \approx \dfrac{1}{c} \mu_p H$

Expressed in terms of e, m, m_n, m_p: $\Theta_p \approx \dfrac{1}{c} \dfrac{e}{m_p} \tau H$

Transition $p \rightarrow n$: $\Theta_n \approx \dfrac{1}{c} \dfrac{(-e)}{(+m_n)} \tau H = -\dfrac{1}{c} \dfrac{e}{m_n} \tau H$

There is a change of sign; therefore, it is possible to distinguish between p-type and n-type.

3. *Tolman experiment.*

Experimental result: $\mathbf{i} = \dfrac{m}{e} \sigma \mathbf{g}$

Expressed in terms of e, m, m_n, m_p: $\mathbf{i} = \dfrac{m}{e} e\mu_p p\mathbf{g} = mp \dfrac{e}{m_p} \tau\mathbf{g}$

Transition $p \rightarrow n$: $\mathbf{i} = (-m)n \dfrac{(-e)}{(+m_n)} \tau\mathbf{g}$

$= mn \dfrac{e}{m_n} \tau\mathbf{g}$

There is no change of sign; therefore, it is not possible to distinguish between p-type and n-type.

Although C. G. Darwin[1] has pointed out that one cannot learn anything concerning the effective mass of the electrons from the Tolman experiment, the opposite has been assumed on occasions.[2] In order to clarify this situation, we pass beyond the somewhat formal substitu-

[1] C. G. Darwin, *Proc. Roy. Soc. (London)*, **A154**: 61 (1936).
[2] Sheldon Brown and S. J. Barnett, *Phys. Rev.*, **87**: 601 (1952).

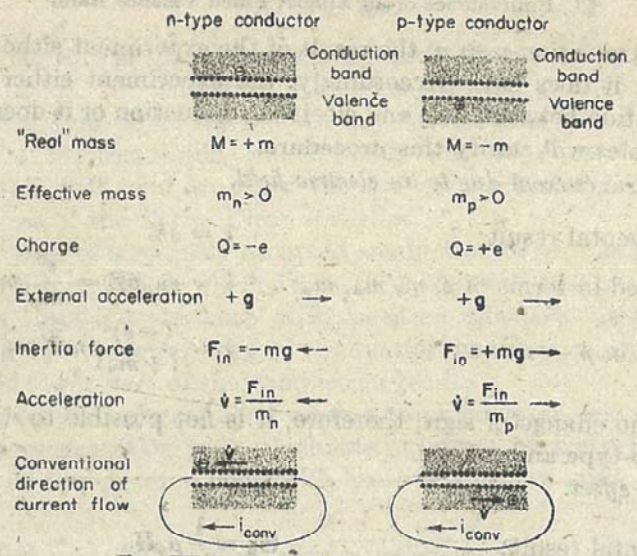

FIG. III.3.2. The Tolman experiment.

tion method of page 67 and represent the experiment with an electron and hole conductor by Fig. III.3.2.

§4. The Failure of the Hole Concept in Problems Involving Electron Interaction

The statements in §3 might give the impression that the hole concept is not only particularly useful in the case of almost filled bands, but also possible and completely permissible in all cases. Concerning the correctness of the last statement we find, upon critical examination of the derivations in §3, that this is not true because of the interaction between electrons. The first step in §3 consisted of the substitution of a filled band with N electrons and of M fictitious positive charges with negative masses for the band with $(N - M)$ electrons and M empty states. The statement, which was made in the course of this first step, that $N - M$ original electrons move in the same manner in both cases is probably correct even if we consider the interaction between the electrons.[1]

In the second step of the foregoing discussion it is stated that in

[1] Incidentally, a closer inspection of the interaction will give an opportunity to clarify the word "fictitious" further. It was chosen to indicate that each of these positive charges does not interact with the "accompanied" electron; for if it existed the interaction would be infinitely large.

transport problems the filled band plus the M additional holes are equivalent to the M fictitious positive charges *alone*, i.e., that the contributions of the N electrons cancel each other. This statement cannot be maintained if interaction between electrons is taken into consideration. In the *presence* of the M fictitious positive charges, the motion of the N electrons of the filled band differs from that in the absence of the M fictitious positive charges. It is, however, only in the latter case that their contributions cancel exactly in transport problems, and we see that the second step in the foregoing discussion, namely, the omission of the electron contribution, is not permissible if interaction is considered.

The band model, however, considers the interaction of electrons only very inadequately, namely, only as a contribution to the fixed lattice potential which is independent of the position of the electron under consideration. The hole representation is equivalent to the electron representation only if the action of the other electrons on the electron under consideration can be represented by such a fixed contribution to the potential. This equivalence was demonstrated in §3 for fixed electric and magnetic fields which are *in*dependent of the electron under consideration. One can state in summary: In so far as an adequate electron representation of processes in solids is possible within the scope of the band model, the hole representation is fully equivalent and in many cases more convenient. Where the interaction between electrons breaks down the validity of the band model, the hole representation also loses its validity. It remains to be seen whether this situation is final, or whether more comprehensive investigations are able to reestablish the hole representation in interaction cases.[1]

§5. Problems (Hall Effect)

1. Give numerical values for the Hall angle for n-type and p-type germanium for H = 1,000 oersteds. What is the transverse voltage across a germanium bar of 1 cm width for a current density of 1 amp/cm^2 and for an impurity density of 10^{15} donors or acceptors/cm^3?

2.* When both electrons and holes are present in a semiconductor traversed by a current, a magnetic field will deflect them to the same side of the crystal, as shown in Fig. III.1.1. If the initial deflection currents were equally large for both types of carriers, no electric charge and therefore no transverse electric field would build up. Generally, however, the initial deflection currents will be different from each other. An electric field will then build up such that the stronger one of the two deflection currents will be retarded while the weaker one, which consists of the opposite charge carriers, will be accelerated. The final field is determined by the

[1] In this connection, see the experiments of K. G. McKay and K. B. McAfee, *Phys. Rev.*, **91**: 1079 (1954); and possibly also E. Spenke, "Hamburger Vorträge," 1954, published by E. Bagge and H. Brüche, Physikverlag Mosbach; in preparation.

condition that the net *electrical* current is zero; this means that the electron and hole *particle* currents are finite and equal.

Show that the Hall-angle Θ for the composite Hall effect is given by

$$\tan \Theta = \frac{p\mu_p^2 - n\mu_n^2}{p\mu_p + n\mu_n} \frac{H}{c} \qquad \text{(III.5.01)}$$

3. What is the value of $\tan \Theta$ according to Eq. (III.5.01) for (*a*) an intrinsic semiconductor, (*b*) a semiconductor with minimal conductivity [see Eqs. (I.3.07) and (I.3.08)]? Give numerical values for germanium for H = 1,000 oersteds. At what conductivity is the Hall angle equal to zero?

Make a qualitative plot of both σ and $\tan \Theta$ as a function of $\log n/n_i$ and show the relationships of the significant points on these curves (assuming $\mu_p < \mu_n$).

4. Calculate the density of the transverse particle current of electron-hole pairs. For what carrier density and what conductivity does this current have a maximum? What is the ratio of the maximal transverse current density to the longitudinal particle current density in germanium for H = 1,000 oersteds?

CHAPTER IV

The Mechanism of Crystal Rectifiers

§1. Introduction

Our ideas about the structure and the mechanism of commercial crystal rectifiers have changed during recent years. The introduction of a radically new type of rectifier is partially responsible for this change: This is the so-called *p-n* junction. It was developed by W. Shockley[1] and a large staff of coworkers at the Bell Telephone Laboratories in the United States. One realization[2] of this new rectifier type consists of a single-crystal germanium wafer (see Fig. IV.1.1). One electrical connection is made by soldering antimony (Sb) to one side and the other by soldering indium (In) to the opposite side. The crystal is subsequently heated to an elevated temperature for some time, so that the antimony and the indium diffuse[3] from their respective sides into the wafer. The transition region, thus introduced, between antimony-containing and indium-containing germanium or *n*-type and *p*-type germanium[4] exhibits pronounced rectifying properties. The voltage-current characteristics (see Fig. IV.1.2) show that large currents with small voltage drops can flow from the indium to the antimony side, whereas in the opposite direction there is only a small residual current even for large voltages.

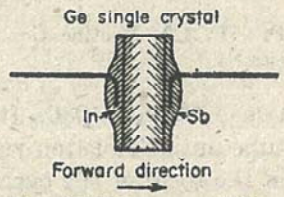

FIG. IV.1.1. *p-n* rectifier. In-Ge-Sb type.

[1] W. Shockley, *Bell System Tech. J.*, **28**: 435 (1949); "Electrons and Holes in Semiconductors," D. Van Nostrand Company, Inc., Princeton, N.J., 1950.

[2] R. N. Hall and W. C. Dunlap, *Phys. Rev.*, **80**: 467 (1950). R. N. Hall, *Proc. IRE*, **40**: 1512 (1952).

[3] It was recently found that alloying processes and subsequent recrystallization, rather than diffusion processes, are responsible for the formation of *p-n* junctions by the method just described. In this connection, see R. R. Law, C. W. Mueller, J. I. Pankove, and L. D. Armstrong, *Proc. IRE*, **40**: 1352 (1952). Similar principles apply to the process of preparing Si *p-n* rectifiers. See G. L. Pearson and B. Sawyer, *Proc. IRE*, **40**: 1348 (1952).

[4] See in this connection pp. 18 to 24.

How did the earlier rectifier types compare? The selenium rectifier of E. Presser,[1] which still dominates the high-current field, consists of a base electrode of iron (Fe) or aluminum (Al) (see Fig. IV.1.3). A crystalline selenium (Se) layer is deposited on this base which, in turn, is covered with an electrode of tin-cadmium alloy (Sn-Cd). The forward current direction is from the base to the cover electrode.

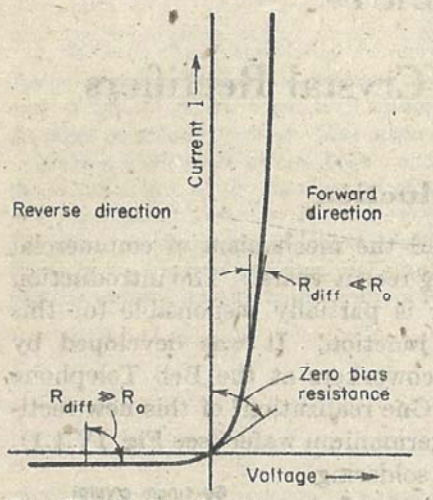

FIG. IV.1.2. Rectification characteristics.

The predecessor of the selenium rectifier is the cuprous oxide rectifier of L. O. Grondahl[2] which consists of an oxidized copper (Cu) plate (see Fig. IV.1.4). The cuprous oxide layer is covered with a graphite or silver electrode. The forward current direction is from the graphite or silver electrode to the copper metal.

The crystal detector discovered by the Strasbourg physics professor Ferdinand Braun[3] in 1874 was widely used in the radio field between 1920 and 1930. It was then almost entirely replaced by the vacuum tube until about ten years later, when it made a sensational comeback in the form of the germanium and silicon diodes for microwave work. In these rectifiers (see Fig. IV.1.5), a metal point makes contact with a germanium (Ge) or silicon (Si) crystal under slight spring pressure.

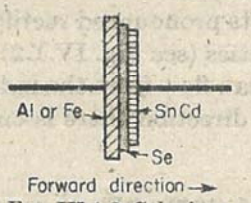

FIG. IV.1.3. Selenium rectifier.

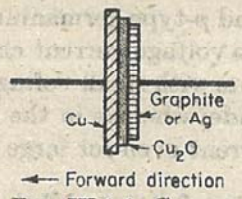

FIG. IV.1.4. Copper oxide rectifier.

The crystal is attached through a nonrectifying large-area contact to its metal holder. In germanium diodes the forward current flows when

[1] E. Presser, *Funkbastler*, p. 558 (1925); *Elektrotech Z.*, **53**: 339 (1932).

[2] L. O. Grondahl, *Science*, **36**: 306 (1926); *J. Am. Insl. Elec. Engrs.*, **46**: 215 (1927).

[3] F. Braun, *Pogg. Ann.*, **153**: 556 (1874); *Wied. Ann.*, **1**: 95 (1877); **4**: 476 (1878); **19**: 340 (1883).

the metal point is positive, in silicon diodes, conversely, when it is negative.

The physical processes which effect the rectification are not distributed over the entire volume of the crystalline semiconductor material. It is known that these processes take place entirely within a "barrier layer" of 10^{-4} to 10^{-5} cm thickness.[1] The barrier layer in selenium rectifiers is at the surface of the selenium near the tin-cadmium alloy; in cuprous oxide rectifiers it is at the boundary of the copper base and in the detectors next to the metal boundary at the point contact.

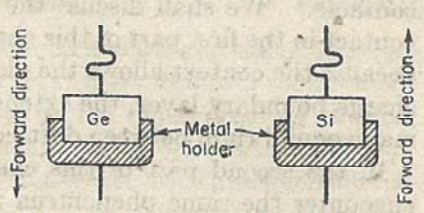

FIG. IV.1.5. Germanium and silicon detectors.

Thus it seemed that the unipolarity of these rectifiers was a consequence of an interaction of *metal and semiconductor* and that the *p-n* junction, with its rectifier effect between antimony and indium-containing germanium, constituted something entirely new. However, B. Davydov[2] pointed out as early as 1938 that strong unipolar effects are to be expected at the junction between *p*-type and *n*-type semiconductors. As a matter of fact, it was shown in recent years that in the selenium rectifier just described the barrier layer does not, in reality, occur at the boundary between the tin-cadmium alloy and the selenium. A chemical reaction between the tin-cadmium electrode and the selenium during the preparation of these rectifiers leads to the formation of a cadmium selenide layer[3] (see Fig. IV.1.6). Poganski[4] as well as Hoffmann and Rose[5] were able to demonstrate beyond doubt that the barrier layer is located at the boundary between the two semiconductors cadmium selenide and selenium, rather than at the boundary between the metal tin-cadmium, and the semiconductor cadmium selenide. A similar situation is most probably realized in germa-

FIG. IV.1.6. CdSe layer in the selenium rectifier. The CdSe layer is deliberately shown too thick for illustrative purposes.

[1] W. Schottky and W. Deutschmann, *Physik. Z.*, **30**: 839 (1929).

[2] B. Davydov, *Tech. Phys. U.S.S.R.*, **5**: 87–95 (1938).

[3] See for instance, Uno Lamm, *ASEA Journal*, **16**: 114 (1939). W. Koch and S. Poganski, *FIAT Rev. Ger. Sci., Final Report* 706, p. 18 (1946). M. Tomura, *Bull. Chem. Soc. Japan*, **22**: 82 (1949); *J. Phys. Soc. Japan*, **5**: 349 (1950). S. Poganski, Dissertation, Technical University Berlin, 1949; *Elektrotech Z.*, **72**: 533 (1951); *Z. Elektrochem.*, **56**: 193 (1952).

[4] S. Poganski, *Z. Physik*, **134**: 469 (1953).

[5] A. Hoffmann and F. Rose, *Z. Physik*, **136**: 152 (1953).

nium and silicon point-contact detectors as well. In any case, special surface treatments and "forming processes"[1] are necessary to obtain good rectifier characteristics, which shows clearly that point-contact detectors do not consist of simple metal-semiconductor contacts.[2]

Nevertheless, rectification effects can occur in metal-semiconductor contacts.[3] We shall discuss the physics of the metal-semiconductor contact in the first part of this chapter, namely, in §1 to §4, principally because the context allows the discussion in simple terms of the space-charge boundary layer, the extension of the carrier density, the Boltzmann equilibrium between diffusion and field current, and other topics.

In the second part of this chapter, namely, in §5 to §8, we shall encounter the same phenomena in the rectification effects at a semiconductor-semiconductor boundary, though in more complex form. This underlines the present-day importance of the original boundary-layer theory of Schottky.[3]

PART 1. THE METAL-SEMICONDUCTOR CONTACT

§2. The Zero-current Condition of a Metal-Semiconductor Contact

The decisive advance of Schottky's boundary-layer theory for crystal rectifiers over all previous theories lies in the realization that special concentration and potential conditions prevail in the boundary region—in the "boundary layer" of the semiconductor—and that these conditions depend on the current through the rectifier.

Let us visualize, for instance, an n-type semiconductor with a uniform donor concentration n_D throughout (see Figs. IV.2.1 and IV.2.2). The concentration n_D of these donors and their dissociation energy E_{cD} shall be so small[4] that all donors have given up their electron $\ominus$ (saturated impurities) at normal temperatures. The concentration n_{D^+} of the positively charged donors is then a constant independent

[1] These are electrical overloads with forward or reverse currents or also alternating current. In spite of certain hypotheses, the forming processes as well as the surface treatments are still essentially an empirical art, for the physics of the point contact is not at all understood.

[2] R. Thedieck, *Physik. Verhandl.*, **3**: 31 (1952); **3**: 212 (1952); *Z. angew. Phys.*, **5**: 165 (1953). L. B. Valdes, *Proc. IRE*, **40**: 445 (1952).

[3] W. Schottky, *Naturwiss.*, **26**: 843 (1938); *Z. Physik*, **113**: 367 (1939); and **118**: 539 (1942). The experimental investigation of rectification effects in metal-semiconductor contacts without an intermediate layer is due to S. Poganski, *Z. Physik*, **134**: 469 (1953).

[4] Concerning the condition for saturated impurities, see pp. 47 to 50.

of location in space

$$n_{D^+} = n_D = \text{const} \tag{IV.2.01}$$

and in the interior of the semiconductor the neutrality condition for the electron concentration n requires

$$n = n_S = n_{D^+} \tag{IV.2.02}$$

However, at the boundary between semiconductor and metal the electron concentration n is determined by an entirely different requirement than the neutrality condition, namely, by the requirement of thermal equilibrium with the metal. For the time being we consider the condition of zero current, so that the number of electrons passing from left

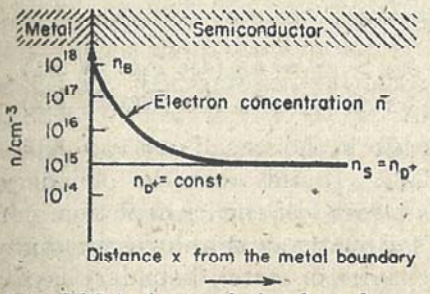

FIG. IV.2.1. Accumulation boundary layer

$(n_B > n_S = n_{D^+})$

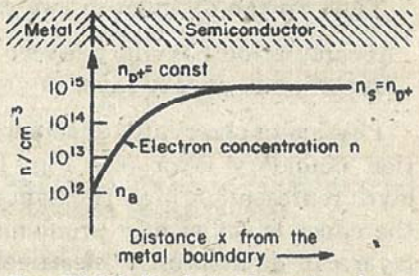

FIG. IV.2.2. Depletion boundary layer

$(n_B < n_S = n_{D^+})$

to right through the metal-semiconductor boundary per unit time and area, because of their random thermal motion, must be equal to the number of electrons passing in the opposite direction. This requirement demands a very specific electron concentration n_B at the semiconductor boundary at a given temperature, and we shall state without further proof[1]

$$n_B = N_C \cdot \mathrm{e}^{-\frac{\Psi_{\text{met sem}}}{kT}} \tag{IV.2.03}$$

where

$$N_C = 2 \cdot \left(\frac{2\pi m_{eff} kT}{h^2}\right)^{3/2} = 2.5 \cdot 10^{19} \left(\frac{m_{eff}}{m}\right)^{3/2} \left(\frac{T}{300°\mathrm{K}}\right)^{3/2} \mathrm{cm}^{-3} \tag{IV.2.04}$$

is an effective density of states in the conduction band. $\Psi_{\text{met sem}}$ is a type of work function of the metal electrons, but refers to electrons passing from the metal into a semiconductor rather than into vacuum as in the usual work function definition.

[1] The reader will find a proof in Eq. (X.7.01). N_C is introduced in Chap. VIII, §1 and §4, see Eqs. (VIII.1.04) and (VIII.4.04).

The most interesting feature of Eq. (IV.2.03) is, for the present, that n_B is determined by the work function $\Psi_{\text{met sem}}$, whereas the concentration n_S in the interior of the semiconductor is determined by the concentration n_{D^+} [see Eq. (IV.2.02)]. The work function $\Psi_{\text{met sem}}$ and the donor density $n_D = n_{D^+}$ are, however, entirely independent of each other, hence also the quantities n_B and n_S.

This fact is not surprising; for the density n_S deep in the interior of the semiconductor must be independent of the composition of the remote metal electrode—Cu or Sn for instance. The boundary density n_B, however, is very strongly affected by the neighboring metal. Thus we see that, in general, n_B and n_S have different values and that $n_B = n_S$ would be a most improbable coincidence.

Thus we can distinguish two cases:

Accumulation boundary layer: $n_B > n_S = n_D$ (see Fig. IV.2.1)
Depletion boundary layer: $n_B < n_S = n_D$ (see Fig. IV.2.2)

Physically observable effects will occur in the second case of a depletion boundary layer (see Fig. IV.2.2). In this case the boundary layer represents a high resistance as a result of carrier depletion. If the effect is sufficiently pronounced, a relatively thin high-resistance layer can dominate the electrical behavior of metal, boundary layer, and semiconductor connected in series. The *reduction* of the resistance of a thin layer, in the opposite case of the accumulation boundary layer, has a negligible effect on the total resistance in view of the constant and much larger resistance of the entire semiconductor body. Accumulation layers are, therefore, of interest only with reference to *nonrectifying* contacts to semiconductors, whereas the metal-semiconductor contact with a depletion boundary layer exhibits rectification properties.

Before we demonstrate this by considering the behavior in the presence of current, we have to discuss the distribution of the electrostatic macropotential[1] V within the boundary layer. This layer is no longer neutral like the interior of the semiconductor, for with $n < n_{D^+}$ there are not enough negative electrons $\ominus$ available to compensate the charge of the positive donors D^+. The boundary layer contains, therefore, a positive space-charge density $\rho(x)$ which, according to the Poisson equation

$$V''(x) = -\frac{4\pi}{\varepsilon}\rho(x) \qquad \text{(IV.2.05)}$$

bends the potential downward. A so-called "diffusion voltage" V_D is present at the boundary layer (see Fig. IV.2.3).

[1] See p. 337 concerning this term.

This potential difference within a conductor in the absence of current is very difficult to comprehend, as experience has shown. First we recall that absence of current does not necessarily exclude potential differences. For instance, the so-called Galvani voltage[1] is established between the *inner* potentials of two conductors of different material in the absence of current; in the same case the so-called Volta or contact potential[1] exists between the *surface* potentials, as is well known from the calculation of the effective grid bias in a vacuum tube. The reader who is interested in more detail concerning this subject and its connection with the diffusion voltage V_D is referred to Chap. X.

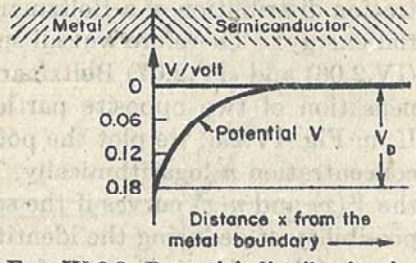

FIG. IV.2.3. Potential distribution in a depletion boundary layer.

It is, however, useful to realize that the diffusion voltage V_D is necessary to establish the zero-current condition. The large concentration gradient from n_S to n_B must lead to a large electron current from right to left. The zero-current condition is established solely by the electrical potential difference V_D which transports the negatively charged electrons from left to right in order to compensate the current resulting from the concentration gradient.

The stratification of the earth's atmosphere is a well-known analogy. Here, too, the absence of vertical motion of the air may be regarded as a compensation of two opposite air currents, namely, the rising air current from the lower high-pressure layers to the higher low-pressure layers and a downward air current due to the gravitational attraction of the earth. The resulting air pressure or the proportional concentration n of the air molecules obeys the so-called barometric equation:

$$n(x) = n(0)\, e^{-\frac{mgx}{kT}} \qquad \text{(IV.2.06)}$$

where m = mass of a molecule
g = gravitational acceleration
k = Boltzmann constant
T = absolute temperature
x = altitude coordinate

If we substitute the electrostatic energy $(-e) \cdot V(x)$ of an electron $\ominus$ in the electrostatic potential $V(x)$ for the potential energy mgx of a molecule in the gravitational field of the earth, we obtain according to (IV.2.06)

[1] See pp. 342 to 345 concerning this term.

$$n(x) = n_S \, e^{+\frac{eV(x)}{kT}} \tag{IV.2.07}$$

In particular, we find at the semiconductor boundary $x = 0$

$$V(0) = -V_D \tag{IV.2.08}$$

and

$$n(0) = n_B \tag{IV.2.09}$$

so that we arrive at

$$n_B = n_S \, e^{-\frac{eV_D}{kT}} \tag{IV.2.10}$$

The relations (IV.2.06) and (IV.2.07) are special cases of the concentration distribution of a Boltzmann gas in a space with varying potential energy. Therefore we call spatial concentration distributions like (IV.2.06) and (IV.2.07) Boltzmann distributions and the mutual compensation of two opposite particle currents Boltzmann equilibrium. If, in Fig. IV.2.3, we plot the potential linearly and in Fig. IV.2.2 the concentration n logarithmically, Eq. (IV.2.07) leads to an identity of the $V(x)$ and $n(x)$ curves if the scales are chosen appropriately. The possibility of deducing the identity of the concentration and potential curves (plotted as indicated) from the existence of a Boltzmann equilibrium is often useful.

This concludes the discussion of the zero-current case. We shall see in §3 what currents arise if the potential difference V_D of the zero-current condition is changed by an externally applied voltage U.

§3. The Metal-Semiconductor Contact with Current

We begin with the assumption of an increase of the potential of the metal electrode by the voltage U_{forw}, while the potential on the semiconductor side is fixed at the far right (see Fig. IV.3.3). In this case the potential drop across the boundary layer is not V_D but only $V_D - U_{\text{forw}}$.

Actually one must distinguish between a voltage U^*_{forw} between the terminals of the rectifier and the portion U_{forw} which lies across the boundary layer itself and is smaller than U^*_{forw}. The neutral portion of the semiconductor adjacent to the boundary layer of Fig. IV.3.2 has also a finite resistance which we shall call the "base resistance R_B." The current through the base causes a voltage drop $R_B I$ which is in series with the boundary-layer voltage U_{forw}:

$$U^*_{\text{forw}} = U_{\text{forw}} + R_B I$$

From the current-voltage characteristics of the boundary layer

$$I = f(U_{\text{forw}})$$

one obtains the over-all characteristic $I = g(U^*_{forw})$ of the entire rectifier including the base resistance R_B (see Fig. IV.3.1). In the following we shall deal only with the characteristics $I = f(U_{forw})$ of the boundary layer.

In order to establish the reduced potential barrier $V_D - U_{forw}$, the space charge which is responsible must also be reduced. The density ρ of this space charge is given very nearly by the fixed concentration n_{D^+} of the immobile donors D^+, for the variable and therefore adjustable concentration n of the mobile electrons $\ominus$ does not play any part in comparison with n_{D^+} in the essential portion of the space-charge boundary layer. A reduction of the space charge, responsible for the potential drop, is possible only through a reduction in the width of the total space charge. The concentration n_S corresponding to neutrality must, therefore, be maintained further to the left than in the zero-current condition (see Fig. IV.3.2). One can visualize this in the form of a "conceptual aid"[1] by assuming

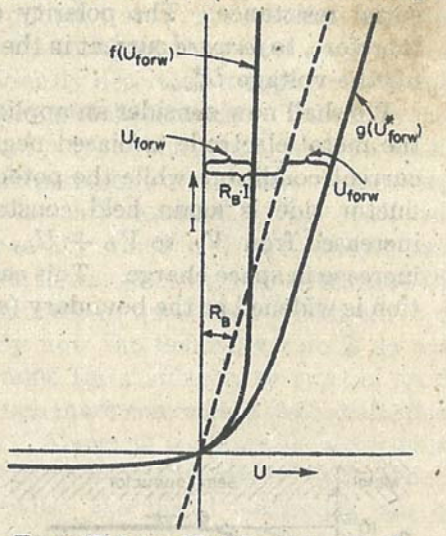

FIG. IV.3.1. Modification of the boundary-layer rectification characteristics $I = f(U_f)$ by the base resistance R_B.

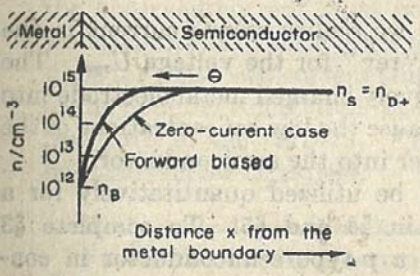

FIG. IV.3.2. Electron concentration n with forward bias.

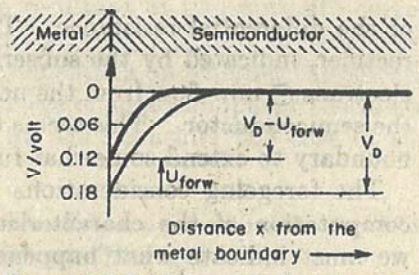

FIG. IV.3.3. Potential distribution $V(x)$ with forward bias.

that the electrons which flow from the interior of the semiconductor toward the positive electrode extend their concentration n_S somewhat further into the boundary zone which, in the absence of current, is depleted of carriers. In any case, the carrier depletion is less intensive

[1] *Only* a "conceptual aid" and not a physically correct picture!

with the assumed bias than in the unbiased condition, and it is certainly plausible that this is accompanied by a reduction of the differential resistance. The polarity of the applied voltage corresponds, therefore, to *forward* current in the rectifier; hence the subscript "forw" for the voltage U^*_{forw}.

We shall now consider an applied voltage of opposite polarity where the metal electrode is biased negatively by U_{rev} relative to the zero-current condition, while the potential far to the right on the semiconductor side is again held constant. The potential barrier is thus increased from V_D to $V_D + U_{rev}$ (see Fig. IV.3.5), which requires an increase in space charge. This can take place only if the carrier depletion is widened at the boundary (see Fig. IV.3.4) leading to an increase

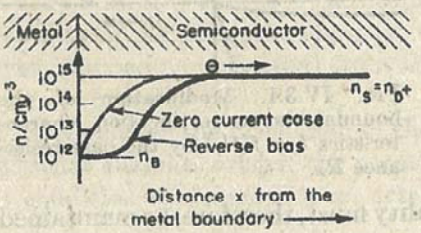

FIG. IV.3.4. Electron concentration n with reverse bias.

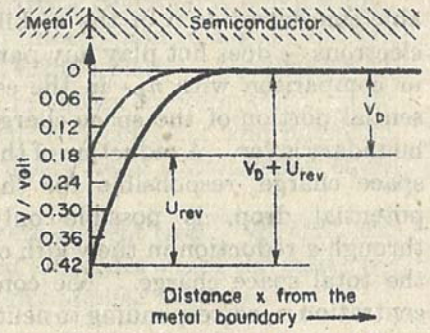

FIG. IV.3.5. Potential distribution $V(x)$ with reverse bias.

of the differential resistance. Thus we realize *reverse* current in the rectifier, indicated by the subscript "rev" for the voltage U_{rev}. The electrons ⊖ now flow from the negatively charged metal electrode into the semiconductor. This seems to cause the low concentrations of the boundary to extend somewhat further into the semiconductor.

The foregoing considerations will be utilized quantitatively for a computation of the characteristics in §4 and §5. To complete §3 we must indicate what happens in a p-type semiconductor in contrast to the previously treated n-type semiconductor. Here again we find that the metal contact effects a change of the hole concentration from the value p_s in the interior of the semiconductor which satisfies the neutrality condition. Again, only the depletion and not the accumulation boundary layers can be observed in view of the series connection of the "base resistance" of the entire semiconductor. The resistance of such layers is dependent on the load; thus, as in the n-type case, the depletion-layer width is reduced along with the differ-

ential rectifier resistance when the positive holes ⊕ are flowing from the interior of the semiconductor toward the metal electrode. For this purpose a negative bias must be applied, which indicates that a p-type semiconductor requires a negative polarity of the metal electrode for forward current, whereas the previously described n-type semiconductor requires a positive polarity of the metal electrode. This difference of the polarities between n-type and p-type conductors, which is easily remembered with the concept of the spreading of the carrier concentration, was predicted by Schottky[1] as early as 1935. This prediction has often been checked experimentally since and has always been found correct. In order to do this it was necessary to determine the conductivity type of the semiconductor in question by Hall effect or thermoelectric measurements. By now the Schottky rule is so well accepted that it is used to determine the conductivity type of semiconductors, since it is generally much more convenient than Hall effect and thermoelectric measurements. Above all it allows the determination of local transitions from p-type to n-type conductivity in a nonhomogeneous semiconductor, because the metal electrodes can be spring-loaded *point* contacts which allow very high resolution.

§4. Calculation of Characteristics

We have seen in the zero-current case that the absence of current can be regarded as a consequence of the exact compensation of two currents in opposite directions. Accordingly, even in the biased case the observed current must be the resultant of two opposite currents which compensate each other at least partially. The analytical expressions for these two currents differ, depending on whether the thickness of the boundary depletion layer is large or small compared to the mean free path of the carriers. First we shall consider the case of small boundary-layer thickness.

a. Boundary-layer Thickness Small Compared with the Mean Free Path of the Electrons (Diode Theory)

Here decelerations of the electrons by collisions with phonons[2] or with imperfections and impurities within the boundary layer can be neglected. The number of electrons arriving per square centimeter and per second from the interior of the semiconductor at the semiconductor interface $x = l$ of the boundary layer (see Fig. IV.4.1) is

[1] W. Schottky: comment on the paper by Störmer, *Z. tech. Phys.*, 16: 512 (1935).
[2] Concerning this term see pp. 13 to 14 and Fig. I.2.9 as well as p. 256.

equal to the unilateral thermal current density[1] $(1/\sqrt{6\pi})v_{th} \cdot n_S$ for the corresponding electron concentration n_S. The fraction

$$\exp - \left(\frac{e(V_D + U_{rev})}{kT}\right)$$

of these electrons possesses the necessary kinetic energy to pass over the peak $+e(V_D + U_{rev})$ of the potential barrier and can therefore reach the boundary of the semiconductor.[2] We thus find at a plane close to the semiconductor boundary $x = 0$ a particle current density from the right to the left

$$\overleftarrow{s} = \frac{1}{\sqrt{6\pi}} v_{th} \cdot n_S \cdot e^{-\frac{e(V_D + U_{rev})}{kT}} \qquad \text{(IV.4.01)}$$

In contrast to this "retarding field" of the electrons which run up against the rise in potential energy from the interior of the semiconductor, the current of electrons from the semiconductor boundary toward the interior is a "saturation current"; for it is aided rather than

[1] The unilateral thermal current density in a Boltzmann gas of the concentration n_S can be calculated to be

$$s = \int v_x \cdot dn = \int_{v_x=-\infty}^{v_x=0} v_x \cdot n_S \cdot \frac{1}{\sqrt{\pi}} \cdot \exp - \left(\frac{v_x}{\sqrt{\frac{2kT}{m_{eff}}}}\right)^2 \cdot d\left(\frac{v_x}{\sqrt{\frac{2kT}{m_{eff}}}}\right)$$

$$s = \frac{1}{\sqrt{\pi}} n_S \sqrt{\frac{2kT}{m_{eff}}} \int_{u=0}^{\infty} u\, e^{-u^2} du = \frac{1}{\sqrt{\pi}} n_S \sqrt{\frac{2kT}{m_{eff}}} \cdot \frac{1}{2}$$

The "mean thermal velocity v_{th}" is given in this book in accordance with the law of equipartition

$$\frac{m_{eff}}{2} v_{th}^2 = \frac{3}{2} kT$$

as the quantity (pp. 256 and 264)

$$v_{th} = \sqrt{\frac{3kT}{m_{eff}}} \qquad \text{(VII.4.22)}$$

This yields the unilateral thermal current density

$$s = \frac{1}{\sqrt{6\pi}} n_S v_{th}$$

[2] In the computation of the fraction of electrons arriving at the peak, we must integrate not to $v = 0$ as in footnote 1, but rather only to

$$v_x = -\sqrt{\frac{2e(V_D + U_{rev})}{m_{eff}}}$$

obstructed by the potential distribution. This aid, however, cannot increase the current beyond the capacity of its source. The saturation current is

$$\vec{s} = \frac{1}{\sqrt{6\pi}} v_{th} \cdot n_B \qquad \text{(IV.4.02)}$$

because the electron concentration n_B at the semiconductor boundary is the source of the current (IV.4.02).

Equation (IV.4.01) corresponds to the reverse-current case, where according to §3 the electrons flow from the semiconductor boundary to

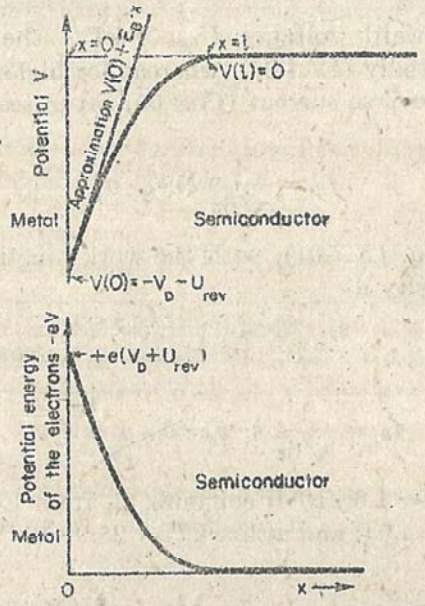

FIG. IV.4.1. The rise of the potential electron energy in a boundary layer. Top: linear approximation of the potential distribution at the metal side of the boundary layer.

the interior of the semiconductor. For the particle current density s, we find by forming the difference $s = \vec{s} - \overleftarrow{s}$

$$s = \frac{1}{\sqrt{6\pi}} v_{th}(n_B - n_S e^{-\frac{e}{kT}V_D} \cdot e^{-\frac{e}{kT}U_{rev}}) \qquad \text{(IV.4.03)}$$

Here we use Eq. (IV.2.10) and obtain

$$s = \frac{1}{\sqrt{6\pi}} v_{th} \cdot n_B(1 - e^{-\frac{e}{kT}U_{rev}}) \qquad \text{(IV.4.04)}$$

Multiplication with the elementary charge e yields a reverse current which flows in conventional terms from right to left:

$$i_{rev} = \frac{1}{\sqrt{6\pi}}\, e \cdot v_{th} \cdot n_B \left(1 - e^{-\frac{e}{kT}U_{rev}}\right) \tag{IV.4.05}$$

or

$$i_{rev} = i_S \left(1 - e^{-\frac{e}{kT}U_{rev}}\right)$$

In the forward direction i_{rev} and U_{rev} are replaced by $-i_{forw}$ and $-U_{forw}$, respectively:

$$i_{forw} = i_S \left(e^{+\frac{e}{kT}U_{forw}} - 1\right) \tag{IV.4.06}$$

For higher forward voltages $U_{forw} \gg kT/e$ the forward current increases exponentially (IV.4.06), whereas for higher reverse voltages $U_{rev} \gg kT/e$ the reverse current (IV.4.05) saturates. The saturation value is

$$i_S = \frac{1}{\sqrt{6\pi}}\, e \cdot v_{th} \cdot n_B \tag{IV.4.07}$$

According to Eq. (X.7.01), with the work function $\Psi_{met\,sem}$ we find the boundary density n_B

$$n_B = N_C\, e^{-\frac{\Psi_{met\,sem}}{kT}} \tag{X.7.01}$$

so that we obtain

$$i_S = \frac{1}{\sqrt{6\pi}}\, e \cdot v_{th} \cdot N_C\, e^{-\frac{\Psi_{met\,sem}}{kT}} \tag{IV.4.08}$$

If we introduce $e = 1.6 \cdot 10^{-19}$ coulomb, v_{th} from Eq. (VII.9.22), and N_C from Eq. (VIII.4.04) and utilize $kT = 25.9 \cdot 10^{-3}$ ev $(T/300°K)$, we arrive at

$$i_S = 1.08 \cdot 10^7 \frac{amp}{cm^2} \left(\frac{m_{eff}}{m}\right) \cdot \left(\frac{T}{300°K}\right)^2 \cdot e^{-38.6\left(\frac{\Psi_{met\,sem}}{e\ volt}\right)\cdot\left(\frac{300°K}{T}\right)}$$

$$= 120 \frac{amp}{cm^2} \cdot \left(\frac{m_{eff}}{m}\right) \cdot \left(\frac{T}{°K}\right)^2 \cdot e^{-\left(\frac{\Psi_{met\,sem}}{e\ volt}\right)\cdot\left(\frac{11{,}580°K}{T}\right)} \tag{IV.4.09}$$

Equations (IV.4.05) and (IV.4.06) give the characteristics for thin boundary layers. We shall now consider the opposite case.

b. Boundary-layer Thickness Large Compared with the Mean Free Path of the Electrons (Diffusion Theory)

Here the electron suffers many collisions with phonons and imperfections or impurities. The electron current from right to left is caused

by the concentration gradient and may be expressed as a diffusion current

$$\overleftarrow{s} = D_n \cdot n'(x) \tag{IV.4.10}$$

(D_n = diffusion constant of the electrons), whereas the current from left to right results from the potential gradient and may be expressed as a field current

$$\overrightarrow{s} = \mu_n n V'(x) \tag{IV.4.11}$$

The diffusion constant D_n and the mobility μ_n of the electrons are connected by the Nernst-Townsend-Einstein relation

$$D_n = \mu_n \cdot \frac{kT}{e} \tag{IV.4.12}$$

so that we find for the total current from left to right, which is source-free and hence independent of location,

$$s = \overrightarrow{s} - \overleftarrow{s} = \mu_n n V'(x) - \mu_n \frac{kT}{e} n'(x) \tag{IV.4.13}$$

According to §3 this particle current direction from the boundary to the interior corresponds to reverse current. We find the *reverse*-current density by multiplication with the elementary charge e

$$i_{rev} = e\mu_n n V'(x) - \mu_n kT n'(x) \tag{IV.4.14}$$

This is a linear differential equation of the first order for the concentration distribution $n(x)$. The solution is

$$n(x) = n_S \, e^{+\frac{e}{kT}V(x)} + \frac{i_{rev}}{\mu_n kT} \int_{\xi=x}^{\xi=l} e^{+\frac{e}{kT}(V(x)-V(\xi))} \, d\xi \tag{IV.4.15}$$

as can be verified by introducing it into (IV.4.14). The only integration constant occurring in the solution of the first-order differential equation has already been determined in (IV.4.15) so that the neutrality concentration n_S of the interior of the semiconductor is obtained at the semiconductor interface $x = l$ of the boundary layer.[1]

If, on the other hand, we apply (IV.4.15) to the metal interface $x = 0$ of the boundary layer, we find according to the qualitative considerations and the figures of §3

$$n(0) = n_B \tag{IV.4.16}$$

[1] The potential value $V(l)$ is made equal to zero at this point (see Fig. IV.4.1).

as well as

$$V(0) = -V_D - U_{rev} \tag{IV.4.17}$$

so that

$$n_B = n_S\, e^{-\frac{e}{kT}V_D}\, e^{-\frac{e}{kT}U_{rev}} + \frac{i_{rev}}{\mu_n kT}\int_{\xi=0}^{\xi=l} e^{+\frac{e}{kT}(V(0)-V(\xi))}\, d\xi \tag{IV.4.18}$$

Using the equation

$$n_S\, e^{-\frac{e}{kT}V_D} = n_B \tag{IV.2.10}$$

we find for the reverse-current density

$$i_{rev} = e\mu_n n_B \cdot \frac{kT}{e} \cdot \frac{1 - e^{+\frac{e}{kT}U_{rev}}}{\int_{\xi=0}^{\xi=l} e^{+\frac{e}{kT}(V(0)-V(\xi))}\, d\xi} \tag{IV.4.19}$$

If the potential function $V(x)$ is known, the integral in the denominator can be evaluated at least in principle, and (IV.4.19) already represents the characteristic $i_{rev} = f(U_{rev})$. The following reasoning often yields an adequate approximation.[1] The exponent $(e/kT)(V(0) - V(\xi))$ of the integral is always negative within the integration range $0 < \xi < l$ (see Fig. IV.4.1). The essential contributions, therefore, occur in the vicinity of $\xi = 0$ where the approximation

$$V(0) - V(\xi) = -\mathbf{E}_B \cdot \xi \tag{IV.4.20}$$

is again applicable. ($\mathbf{E}_B$ = contribution of the boundary field strength, see Fig. IV.4.1.) The integral in the denominator thus becomes

$$\int_{\xi=0}^{\xi=l} e^{\frac{e}{kT}(V(0)-V(\xi))}\, d\xi \approx \int_{\xi=0}^{\xi=l} e^{-\frac{e\mathbf{E}_B}{kT}\xi}\, d\xi = \frac{kT}{e\mathbf{E}_B}\left[e^{-\frac{e\mathbf{E}_B}{kT}\xi}\right]_{\xi=l}^{\xi=0} \tag{IV.4.21}$$

Since $e\mathbf{E}_B l/(kT) \gg 1$ the term corresponding to $\xi = l$ can be omitted. We find thus simply

$$\int_{\xi=0}^{\xi=l} \cdots\, d\xi \approx \frac{kT}{e\mathbf{E}_B} \tag{IV.4.22}$$

and obtain from (IV.4.19) the equation of the characteristic

$$i_{rev} \approx e\mu_n n_B \mathbf{E}_B\,(1 - e^{-\frac{e}{kT}U_{rev}}) \tag{IV.4.23}$$

[1] But not always! The case of unsaturated impurities in the boundary layer cannot be treated accurately in this manner.

For forward current the substitution $i_{rev} \rightarrow -i_{forw}$ and $U_{rev} \rightarrow -U_{forw}$ yields

$$i_{forw} \approx e\mu_n n_B \mathbf{E}_B (e^{+\frac{e}{kT}U_{forw}} - 1) \qquad \text{(IV.4.24)}$$

The comparison with the characteristic equations (IV.4.05) and (IV.4.06) of the diode theory shows the great similarity between the results of the two theories. The saturation current of the boundary concentration n_B

$$i_S = \frac{1}{\sqrt{6\pi}} e v_{th} n_B \qquad \text{(IV.4.07)}$$

is replaced by the field current corresponding to this concentration:

$$i_{field} = e\mu_n n_B \mathbf{E}_B \qquad \text{(IV.4.25)}$$

Unlike the previously mentioned saturation current, i_{field} is dependent also on the applied voltage through the boundary layer field strength $\mathbf{E}_B$. Compared to the exponential term in (IV.4.23) or (IV.4.24), this dependence is significant. The reader will find further particulars in the literature.[1]

§5. The Concentration Distribution in a Boundary Layer

Even with current flow the concentration distribution $n(x)$ follows the Boltzmann law $n(x) = n_S e^{+\frac{eV(x)}{kT}}$ in most of the boundary layer. Only in the immediate proximity of the metal is the concentration increased for reverse current and decreased for forward current (see Fig. IV.5.1) with respect to the Boltzmann distribution. For reverse current the diffusion current is so much reduced that i_{rev} is almost a pure field current in the boundary-layer regions adjoining the metal. For forward current the diffusion current is so much increased that i_{forw} is here almost a pure diffusion current.

Equation (IV.4.15) gave for the concentration distribution:

$$n(x) = n_S e^{+\frac{e}{kT}V(x)} + \frac{i_{rev}}{\mu_n kT} \int_{\xi=x}^{\xi=l} e^{+\frac{e}{kT}(V(x)-V(\xi))} d\xi \qquad \text{(IV.4.15)}$$

This can be simplified by noting the negative sign of the exponent in the integral, as in the calculation of the characteristics, hence deducing

[1] W. Schottky, *Z. Physik*, **118**: 539 (1942). E. Spenke, *Z. Physik*, **126**: 67 (1949); *Z. Naturforsch.*, **4a**: 37 (1949).

that appreciable contributions occur only in the vicinity of $\xi = x$, where we can approximate as follows:

$$V(x) - V(\xi) \approx -\mathbf{E}(x)(\xi - x) \qquad \text{with } \mathbf{E}(x) > 0 \qquad \text{(IV.5.01)}$$

We obtain the integral

$$\int_{\xi=x}^{\xi=l} e^{+\frac{e}{kT}(V(x)-V(\xi))}\, d\xi \approx \frac{kT}{e\mathbf{E}(x)}\left[1 - e^{-\frac{e\mathbf{E}(x)}{kT}(l-x)}\right] \qquad \text{(IV.5.02)}$$

If we forego a description of the concentration conditions at $x \approx l$, we can neglect the exponential term in the parenthesis, and we obtain

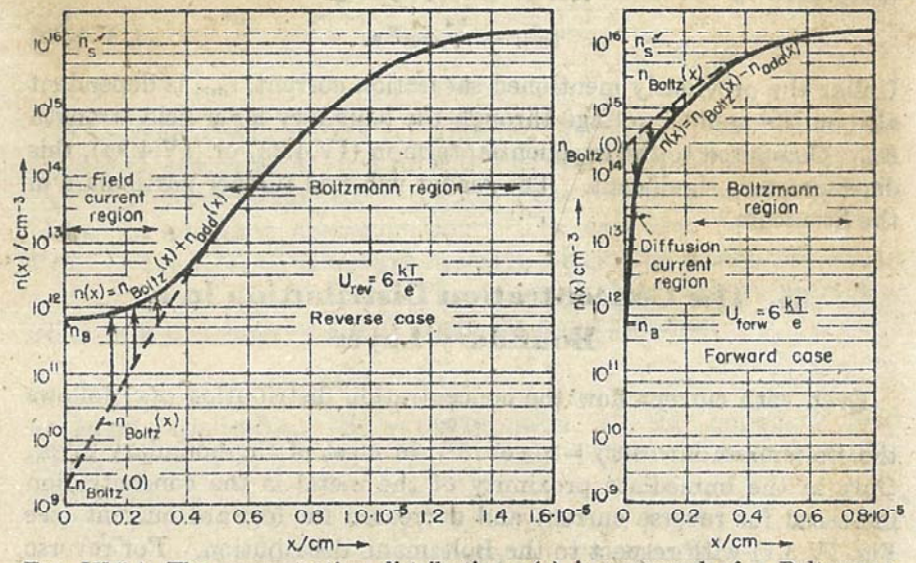

FIG. IV.5.1. The concentration distribution $n(x)$ is composed of a Boltzmann distribution $n_{\text{Boltz}}(x)$ and an added concentration $n_{\text{add}}(x)$. The figure is based on the values $n_S = 1.3.10^{16}$ cm^{-3} and $V_D = 10\frac{kT}{e} = 0.259$ volt. With respect to the additional assumptions and simplifications see E. Spenke, *Z. Physik*, **126**: 67 (1949).

for the concentration distribution in the reverse-current case

$$n(x) \approx n_S\, e^{+\frac{e}{kT}V(x)} + \frac{i_{\text{rev}}}{e\mu_n \mathbf{E}(x)} \qquad \text{for } x < l \qquad \text{(IV.5.03)}$$

For forward current the substitution $i_{\text{rev}} \to -i_{\text{forw}}$ leads to

$$n(x) \approx n_S\, e^{+\frac{e}{kT}V(x)} - \frac{i_{\text{forw}}}{e\mu_n \mathbf{E}(x)} \qquad \text{for } x < l \qquad \text{(IV.5.04)}$$

These equations contain very important physical information. They indicate that the electron distribution $n(x)$ is composed of

two components (see Fig. IV.5.1) of which the first is the Boltzmann distribution

$$n_{\text{Boltz}}(x) = n_S \, e^{+\frac{e}{kT}V(x)} \tag{IV.5.05}$$

and therefore does not correspond to any current, since the field and diffusion current compensate each other exactly.

However, the Boltzmann component $n_{\text{Boltz}}(x)$ has a boundary value which for reverse current is too small by a factor $e^{-\frac{e}{kT}U_{\text{rev}}}$

$$n_{\text{Boltz}}(0) = n_S \, e^{+\frac{e}{kT}V(0)} = n_S \, e^{-\frac{e}{kT}V_D} \, e^{-\frac{e}{kT}U_{\text{rev}}} = n_B \, e^{-\frac{e}{kT}U_{\text{rev}}} \tag{IV.5.06}$$

and for forward current is too large by a factor $e^{+\frac{e}{kT}U_{\text{forw}}}$

$$n_{\text{Boltz}}(0) = n_B \, e^{+\frac{e}{kT}U_{\text{forw}}} \tag{IV.5.07}$$

The correct boundary value n_B is now established if an additional concentration

$$n_{\text{add}}(x) = \frac{|i|}{e\mu_n \mathbf{E}(x)} \tag{IV.5.08}$$

is added for reverse current according to (IV.5.03) and is subtracted for forward current according to (IV.5.04).

This influences not only the concentration $n(x)$ but also the concentration gradient $n'(x)$ (see Fig. IV.5.1). The concentration gradient and with it the diffusion current are tremendously reduced for reverse current when compared to the pure Boltzmann component $n_{\text{Boltz}}(x)$ with its exact compensation of field and diffusion current. For all practical purposes, the field current $i_{\text{field}} = e\mu_n n(x)\mathbf{E}(x)$ alone remains, which becomes simply

$$i_{\text{field}} \approx e\mu_n \cdot n_{\text{add}}\mathbf{E}(x) \tag{IV.5.09}$$

since $n_{\text{Boltz}} \ll n_{\text{add}}$. This leads with (IV.5.08) to

$$i_{\text{field}} \approx i_{\text{rev}} \tag{IV.5.10}$$

For forward current, however, the concentration gradient is increased as compared to the pure Boltzmann contribution (see Fig. IV.5.1). The diffusion current greatly predominates over the field current, and the entire forward current i_{forw} is for all practical purposes diffusion current.

Figure IV.5.1 shows, however, that the influence of n_{add} is noticed only at the metal interface of the boundary layer. Toward the

interior of the semiconductor we find

$$n_{\text{Boltz}} \gg n_{\text{add}}$$

and the distribution is a Boltzmann distribution not only for zero current but also for reverse and forward current. This is again very plausible. Let us take, for instance, reverse current and go from left to right through the boundary layer in Fig. IV.5.1. To the extreme left we find the field-current region, where the density increases at such a rate that, with the decreasing field strength, the required total current is carried as field current. This increase leads to an increase of the concentration gradients.[1] This in turn causes the diffusion current in the opposite direction to increase which, to begin with, is immaterial. However, further to the right the diffusion current in the opposite direction approaches the same order of magnitude as the field current. Both currents must then have become large compared to the total current i_{rev} in order to leave a residual current i_{rev} in spite of the mutual compensation. We have now passed over from the field-current region on the left to the Boltzmann region on the right.

Thus in the region of the boundary layer adjoining the semiconductor, the Boltzmann distribution prevails substantially not only for zero current but also for current flow. Therefore, the logarithmically plotted concentration distribution $n(x)$ and the potential distribution $V(x)$ are identical for current flow as well as for zero current as shown on page 78.

PART 2. THE p-n JUNCTION

§6. The Zero-current Condition of a p-n Junction

The first five paragraphs of this chapter were concerned with the properties of a semiconductor-metal contact. A systematic procedure would require next the treatment of a contact between two different semiconductors. However, the multitude of possible phenomena is

[1] When considering the concentration gradient, it must be pointed out that the concentration is plotted as $\log_{10} n(x)$ and not as $n(x)$. Thus we obtain the concentration gradient dn/dx from the slope $d \ln n/dx$ by multiplying with the concentration $n(x)$ itself:

$$\frac{dn}{dx} = n(x) \cdot \frac{d \ln n}{dx}$$

The slope $d \ln n/dx$ increases somewhat from left to right. However, the factor $n(x)$ increases much more rapidly, so that the concentration gradient dn/dx increases very rapidly from left to right.

here so large that we are forced to make a choice. In view of their practical importance, we shall limit ourselves to the so-called *p-n junctions*. A p-n junction is *not* a contact between two entirely different semiconductors such as selenium and cadmium selenide; the term "junction" is chosen to indicate that only one single host lattice is present and that there is a transition from a zone doped with acceptors producing holes into a zone doped with donors producing electrons.

Referring to the example discussed in §1, we imagine—see Figs. IV.1.1 and IV.6.1—that indium (In) is diffused into a germanium single crystal from the left, creating acceptor impurities A^- and holes $\oplus$. This leads to p-type germanium on the left. Antimony (Sb) is diffused into the crystal from the right, creating donor impurities D^+ and electrons $\ominus$. This leads to n-type germanium on the right. To begin with we consider the special case of a symmetrical p-n junction. Let us assume, for instance, an impurity concentration[1] of 10^{16} cm^{-3}:

$$n_{A^-} = n_{D^+} = 10^{16}\ \text{cm}^{-3} \tag{IV.6.01}$$

The mean distance between two impurities is thus

$$\sqrt[3]{\frac{4.52 \cdot 10^{22}\ \text{cm}^{-3}}{10^{16}\ \text{cm}^{-3}}} = 165$$

interatomic distances in germanium because the concentration of the germanium atoms is $4.52 \cdot 10^{22}$ cm^{-3}. For reasons of neutrality, the germanium on the left must contain a hole concentration

$$p_p = 10^{16}\ \text{cm}^{-3} \tag{IV.6.02}$$

and on the right an electron concentration

$$n_n = 10^{16}\ \text{cm}^{-3} \tag{IV.6.03}$$

In addition, the so-called thermal pair formation (see page 25ff) introduces throughout the germanium pairs of electrons and holes continuously so that the p germanium contains a certain electron concentration n_p and the n germanium contains a certain hole concentration p_n. For zero current these concentrations result from thermal equilibrium between pair formation and its counterprocess, recombination. The electron concentration n and the hole concentration p obey (but only in the case of thermal equilibrium) a law of mass action[2]

$$n \cdot p = n_i^2 \tag{IV.6.04}$$

[1] Except at very low temperatures all the substitutional impurities in germanium at the concentrations considered here are dissociated.

[2] See pp. 26 and 305.

The so-called intrinsic density n_i for germanium at room temperature has approximately the value[1]

$$n_i \approx 10^{13}\ \text{cm}^{-3} \tag{IV.6.05}$$

This results in an electron concentration on the left in the p germanium of

$$n_p = 10^{10}\ \text{cm}^{-3} \tag{IV.6.06}$$

and an equal hole concentration on the right in the n germanium of

$$p_n = 10^{10}\ \text{cm}^{-3} \tag{IV.6.07}$$

From the standpoint of space charge, these 10^6 times smaller concentrations can, of course, be neglected compared with n_n and p_p.

We make the unrealistic but convenient and fundamentally permissible assumption that the impurity concentrations n_{A^-} and n_{D^+} maintain a constant value in space of 10^{16} cm^{-3} on the left and on the right, dropping abruptly to zero at the center $x = 0$. This assumption introduces an abrupt transition from indium doping to antimony doping at the center. Such an abrupt transition is obviously not shared by the concentrations n and p of the *mobile* holes and electrons. The concentrations n and p must exhibit a spatial distribution differing from that of the impurity concentrations n_{A^-} and n_{D^+} because they do not drop from their neutrality values p_p and n_n, respectively, to zero after crossing the center $x = 0$ but rather to the equilibrium values $p_n = n_i^2/n_n$ and $n_p = n_i^2/p_p$, respectively, as required by the mass-action law (IV.6.04). In this §6 we shall discuss the distribution of p and n in the case of thermal equilibrium corresponding to zero current.

The transition from p_p to p_n and from n_n to n_p, respectively, must assume the form of an S-shaped curve (see Fig. IV.6.1). Space charges are formed on the left and on the right of the center $x = 0$. The positive space-charge density in the region $x > 0$ bends the electrostatic potential $V(x)$ downward (see Fig. IV.6.1) according to the Poisson equation

$$V''(x) = -\frac{4\pi}{\varepsilon}\rho(x) < 0 \qquad \text{for } x > 0 \tag{IV.6.08}$$

whereas the negative space-charge density connecting at the left bends the potential $V(x)$ back to horizontal:

$$V''(x) = -\frac{4\pi}{\varepsilon}\rho(x) > 0 \qquad \text{for } x < 0 \tag{IV.6.09}$$

[1] More accurately, $2.5 \cdot 10^{13}$ cm^{-3}. See E. M. Conwell, *Proc. IRE*, 40: 1329 (1952), Table II.

The over-all result is—as for the boundary layer of a metal-semiconductor contact—a potential step[1] within which is established a Boltzmann distribution

$$p(x) = p_n \, e^{-\frac{e}{kT}V(x)} \tag{IV.6.10}$$

$$n(x) = n_n \, e^{+\frac{e}{kT}V(x)} \tag{IV.6.11}$$

The height of the step, i.e., the diffusion voltage V_D, can be calculated

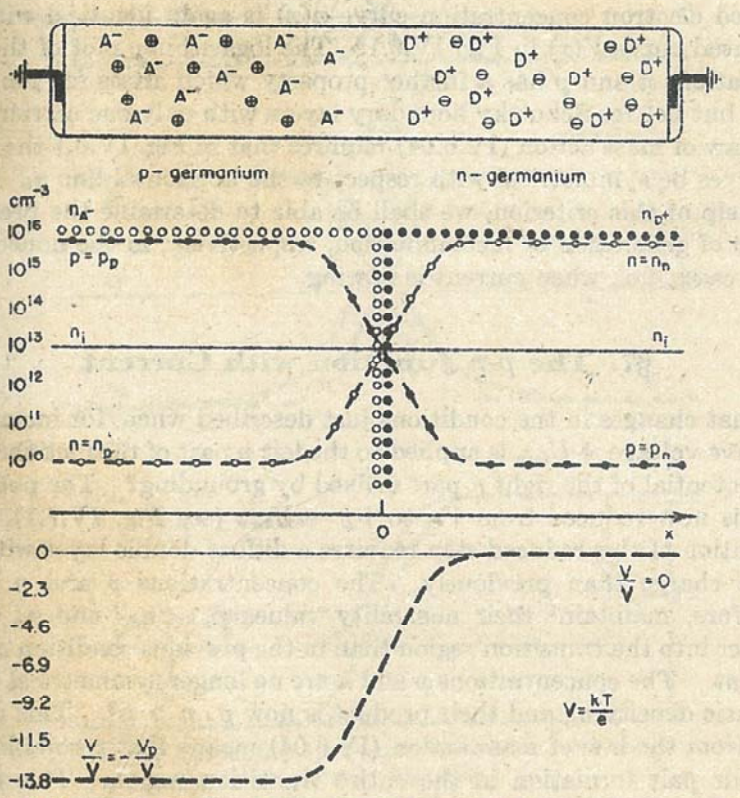

FIG. IV.6.1. Potential distribution and concentration distribution in a p-n junction. Zero-current case.

by solving these equations for $x \to -\infty$ with $V(-\infty) = -V_D$:

$$\frac{p_p}{p_n} = \frac{n_n}{n_p} = e^{+\frac{e}{kT}V_D} \tag{IV.6.12}$$

[1] The potential step is present in spite of the fact that the germanium crystal is grounded at both terminals! See also pp. 76 and 77 and pp. 359 to 361.

In the present case, V_D would be

$$V_D = \frac{kT}{e} \ln \frac{n_n}{n_p}$$
$$\approx 25.9 \text{ mv} \cdot \ln \frac{10^{16}}{10^{10}} = 25.9 \text{ mv} \cdot 13.8 \qquad \text{(IV.6.13)}$$
$$V_D \approx 0.358 \text{ volt}$$

Because of the Boltzmann equilibrium (IV.6.11), the logarithmically plotted electron concentration curve $n(x)$ is again identical with the potential curve $V(x)$ in Fig. IV.6.1. The logarithmic plot of the concentrations n and p has a further property which arises for p-n junctions but not for Schottky boundary layers with only one carrier type. The law of mass action (IV.6.04) requires that in Fig. IV.6.1 the n and p curves be symmetrical with respect to the horizontal line n_i. With the help of this criterion, we shall be able to determine the predominance of generation or recombination, respectively, in the nonequilibrium cases, i.e., when current is flowing.

§7. The p-n Junction with Current

What changes in the conditions just described when, for instance, a positive voltage $+U_{forw}$ is applied to the left p part of the rectifier while the potential of the right n part is fixed by grounding? The potential step is now reduced from V_D to $V_D - U_{forw}$ (see Fig. IV.7.1). The formation of this reduced step requires a diffuse double layer with less space charge than previously. The concentrations p and n must, therefore, maintain[1] their neutrality values $p_p = n_{A^-}$ and $n_n = n_{D^+}$ further into the transition region than in the previous condition of zero current. The concentrations p and n are no longer symmetrical to the intrinsic density n_i, and their product is now $p \cdot n > n_i^2$. This deviation from the law of mass action (IV.6.04) means that recombination exceeds pair formation in the entire transition region. This comes about in the following way: The potential increase at the left end of the rectifier drives the positive holes of the p region from left to right, namely, toward the transition region. The negative electrons of the right n region are attracted by the left positive electrode and thus flow, also, toward the transition region. Therefore the concentrations n and p increase within the transition region. This leads to an increase of the recombination rate $r \cdot n \cdot p$, whereas the generation g remains constant since it is independent of the concentration. A new steady-

[1] This argument should be made more precise as on p. 79.

state condition is established, when the excess of recombination over generation is just compensated by the influx into the transition region from both sides. The foregoing shows furthermore that the total current i_{forw} is a pure hole current i_p on the left side and a pure electron current i_n on the right side (see Fig. IV.7.1, top).

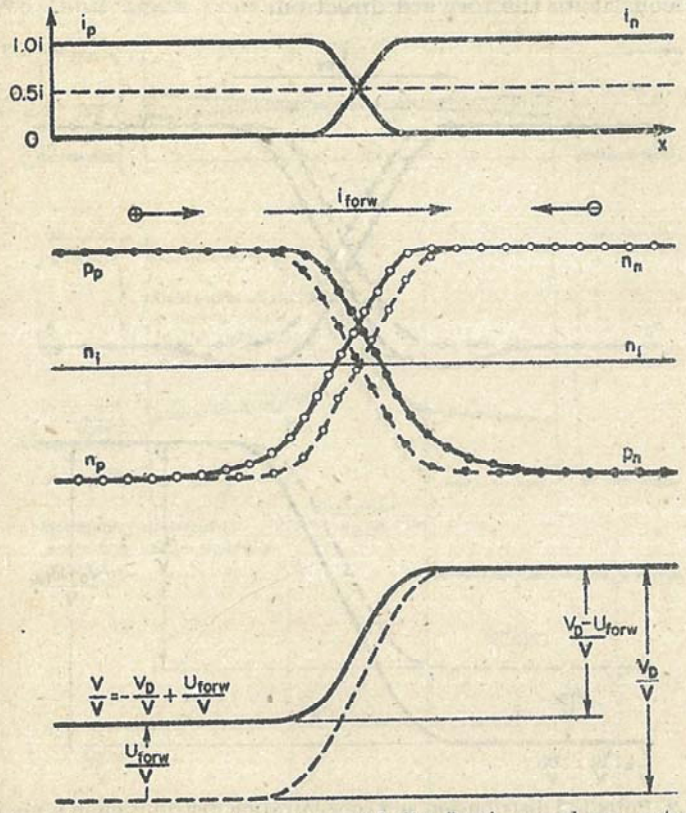

FIG. IV.7.1. Current distribution, potential distribution, and concentration distribution in a p-n junction. The forward bias case.

It becomes now plausible that the experimentally observed high resistance of a p-n rectifier at small bias (the "zero-bias resistance" in Fig. IV.1.2) arises from carrier depletion in the transition region. There the carrier densities p and n drop to the intrinsic density n_i. As we saw at the end of Chap. I (p. 27), this condition of intrinsic conduction corresponds to the highest attainable specific resistance in a semiconductor. This cause of the high zero-bias resistance of the p-n rectifier has been reduced by the application of a positive voltage to the left p end of the rectifier. The high carrier concentrations of the

p part and n part are, as we have seen, effectively carried along by the hole and electron currents directed toward the transition region, so that they are, so to speak, extended into the otherwise depleted transition region. The resulting reduction in resistance of the transition region, and hence of the p-n rectifier, shows that this polarity and the direction of the conventional current from left to right through the rectifier constitute the forward direction.

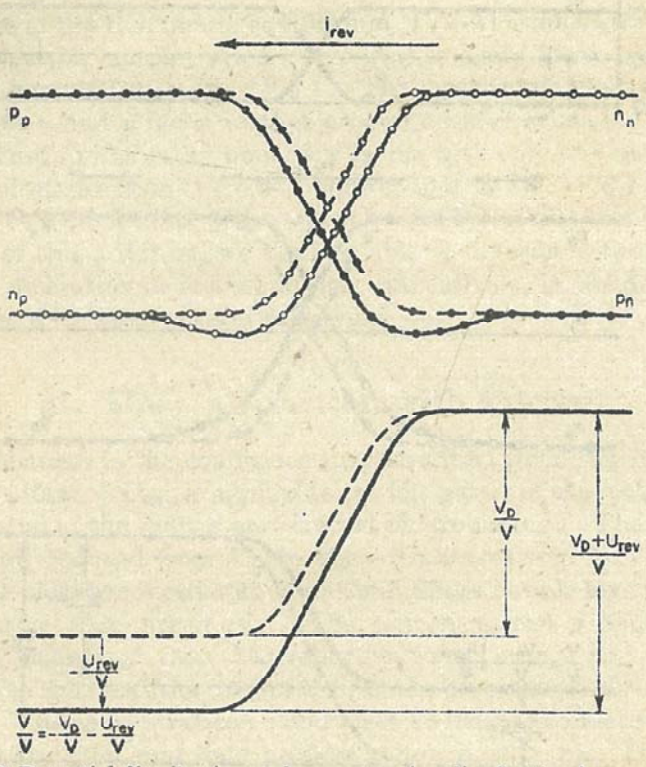

FIG. IV.7.2. Potential distribution and concentration distribution in a p-n junction. The reverse bias case.

The proof that the opposite direction of the conventional current is the reverse direction can now be made quite brief (see Fig. IV.7.2). A negative potential $-U_{rev}$ must be applied to the left end so that the conventional current flows from right to left. The potential step in the transition region increases thus to $V_D + U_{rev}$ and requires for its formation more space charge. This can be attained only by widening the depleted transition region. The widening of the high-resistance transition region increases the rectifier resistance. Thus the reverse direction is realized.

§8. The Special Case of a Low Recombination p-n Junction According to Shockley

Shockley has pointed out that the use of crystals with very low recombination rates leads to very special characteristics of p-n rectifiers. We shall again consider the case of the forward direction,

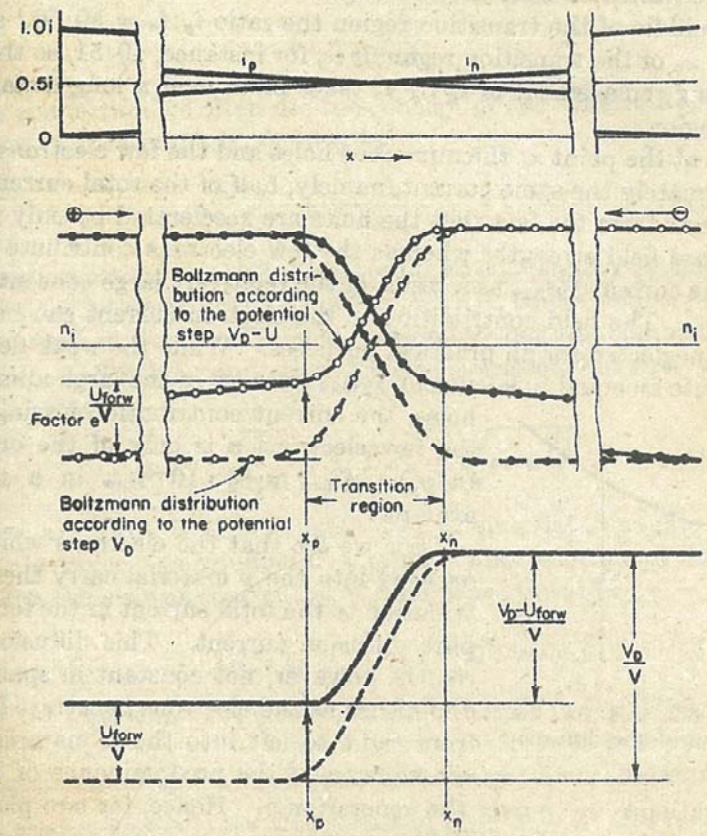

FIG. IV.8.1. p-n junction with low recombination. The forward-bias case. The potential in the forward-bias case is not strictly horizontal (constant) to the left of x_p and to the right of x_n, but rather a slight drop from left to right is present. In these "base regions" weak base fields are present which, however, are too small to be clearly incorporated in the above figure.

namely, the polarity shown in Fig. IV.7.1. We have pointed out previously that the current in the rectifier is carried by a hole current at the far left and by an electron current in the opposite direction at the far right and that the hole current passes over into the oppositely directed electron current in the transition region owing to the predomi-

nance of recombination. If, according to Shockley, special measures, which we shall describe later, are taken to reduce the recombination appreciably, the holes will penetrate deeply into the n region and the electrons into the p region (see Fig. IV.8.1). The $\oplus$ current begins to be taken over by the $\ominus$ current of opposite direction deep within the p region, long before the transition region is reached. At the beginning x_p of the transition zone the exchange is already 49 per cent complete, in the middle of the transition region the ratio $i_p : i_n = 50:50$,[1] and at the end x_n of the transition region it is, for instance, 49:51, so that the remaining replacement of i_p by i_n takes place over a long distance in the n region.

Thus at the point x_p the numerous holes and the few electrons carry approximately the same current, namely, half of the total current i_{forw}. This stems from the fact that the holes are accelerated by only a very weak base field strength[2] whereas the few electrons contribute about the same current $\frac{1}{2}i_{forw}$ as a result of the relatively large concentration gradient. The field contribution to the electron current can be completely neglected for all practical purposes. While the weak field can contribute an equal hole current $\frac{1}{2}i_{forw}$ because of the large number of holes, the current contribution arising from the few electrons n is only of the order of $(n/p_p) \cdot \frac{1}{2}i_{forw} \approx \frac{1}{2} \cdot 10^{-4} i_{forw}$ in a typical example.

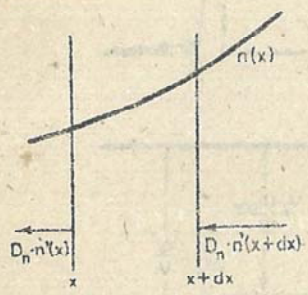

FIG. IV.8.2. Aid for the derivation of the diffusion equation.

Thus we see that the electrons which are dragged into the p material carry their contribution to the total current in the form of a pure diffusion current. This diffusion current is, however, not constant in space, but diminishes as the electrons penetrate further from right to left into the p material as a consequence of the predominance of the recombination $r \cdot n \cdot p$ over the generation g. Hence, for two planes at x and $x + dx$ (see Fig. IV.8.2) we obtain

$$D_n \cdot n'(x + dx) - D_n \cdot n'(x) = (r \cdot n(x) \cdot p_p - g)\, dx \qquad \text{(IV.8.01)}$$

Applying the law of mass action (IV.6.04) to the p germanium, we obtain

$$g = r \cdot n_p \cdot p_p \qquad \text{(IV.8.02)}$$

and using this on the right side of (IV.8.01) we obtain

$$+D_n n''(x)\, dx = r p_p (n(x) - n_p)\, dx$$

[1] We still assume the symmetrical transition $n_n = n_p (= 10^{13}\ \text{cm}^{-3})$.

[2] In this connection, see also the caption to Fig. IV.8.1.

or in another form

$$\frac{d^2}{dx^2}(n(x) - n_p) - \frac{1}{L_n^2}(n(x) - n_p) = 0 \qquad \text{(IV.8.03)}$$

where we have introduced the so-called diffusion length of the electrons in p germanium

$$L_n = \sqrt{\frac{D_n}{rp_p}} \qquad \text{(IV.8.04)}$$

In this connection we often use the concept of the "lifetime τ_n" of the electrons in p germanium, defined by[1]

$$\tau_n = \frac{1}{rp_p} \qquad \text{(IV.8.05)}$$

The justification of this definition would be too lengthy at this point. We refer the interested reader to Chap. IX, §1, where similar problems are discussed in detail. We shall note only that the combination of (IV.4.12), (IV.8.05), and the Nernst-Townsend-Einstein relation (IV.4.12) leads to

$$L_n = \sqrt{\frac{D_n}{rp_p}} = \sqrt{D_n\tau_n} = \sqrt{\mu_n\tau_n\frac{kT}{e}} \qquad \text{(IV.8.06)}$$

L_n and τ_n are added to the conductivity σ as material constants characterizing a given semiconductor sample. The solution of (IV.8.03) which fits the electron current as it diminishes toward $x = -\infty$ is

$$n(x) = n_p + C\,e^{+\frac{x-x_p}{L_n}} \qquad \text{(IV.8.07)}$$

so that the logarithm of the concentration plotted in Fig. IV.8.1 shows a linear decrease toward $-\infty$ as long as $n(x)$ is large compared with the equilibrium value n_p. The slope of this straight line is determined by

[1] This relation applies to the specific reaction $\ominus + \oplus \to 0$. The carrier annihilation does not necessarily have to follow this recombination equation. Atomic imperfections and impurities as well as more extended lattice disorders, such as interfaces and surfaces, play an important role here. However, there is little doubt that in a first approximation the deviation $n - n_p$ from the equilibrium concentration n_p is the determining factor. Hence the expression

$$\frac{1}{\tau_n}(n - n_p)$$

for the number of carriers disappearing per unit time and volume has a rather general significance—independent of the particular recombination mechanism $\ominus + \oplus \to 0$. We shall, therefore, in the following always give the equations in a form which makes use of $\tau_n = 1/rp_p$ since this relation is tied to the specific reaction $\ominus + \oplus \to 0$.

the diffusion length L_n in such a fashion that the concentration $n(x)$ drops by a factor e over a distance L_n.

We shall now calculate the current which is carried by the diffusion tail; from (IV.8.07) we obtain

$$D_n n'(x) = \frac{D_n}{L_n} C\, e^{+\frac{x-x_p}{L_n}} = \frac{D_n}{L_n} (n(x) - n_p) \qquad \text{(IV.8.08)}$$

Thus i_n, the contribution to the total current i_{forw} which is carried by electrons, is at the point $x = x_p$:

$$i_n(x_p) = \frac{eD_n}{L_n} (n(x_p) - n_p) \qquad \text{(IV.8.09)}$$

For the computation of the characteristics, we must determine the concentration increase $n(x_p) - n_p$ at the beginning $x = x_p$ of the electron diffusion tail on the left (see Fig. IV.8.1) as a function of the voltage U_{forw} applied to the p-n junction. This can be done by stating the Shockley condition of "low" recombination more precisely. Low recombination means a small recombination coefficient and therefore, according to (IV.8.05), long lifetime τ_n and, according to (IV.8.04), long diffusion length L_n. The requirement of "low" recombination means actually that the diffusion length L_n is long compared with the width $x_n - x_p$ of the transition region. In this more precise form the Shockley condition has decisive consequences. The electron concentration drops within the small transition region $x_n - x_p$ by several or many powers of e, whereas in the diffusion tail the drop within the large diffusion length L_n is only by a factor e. The concentration gradient and with it the diffusion current must, therefore, increase enormously in the transition from diffusion tail to transition region. This is possible only if the excessive diffusion current in the transition region is compensated by a field current of almost equal magnitude, for the current contribution i_n of the electrons remains practically unchanged. This indicates that an approximate Boltzmann equilibrium exists in the transition region. This, in turn, leads to the potential curve $V(x)$, the logarithmically plotted electron concentration $n(x)$ being identical within the transition region. Since this applies also for zero current, it is evident from Fig. IV.8.1 that the rise by the forward bias U_{forw} of the potential curve $V(x)$ in the p region is associated with a rise in the concentration curve $n(x)$ at the point $x = x_p$ by the factor $e^{+\frac{e}{kT}U_{forw}}$:

$$n(x_p) = n_p\, e^{+\frac{e}{kT}U_{forw}} \qquad \text{(IV.8.10)}$$

This, however, is the relation between the voltage U_{forw} and the con-

centration $n(x_p)$ at the beginning $x = x_p$ of the electronic diffusion tail for which we were looking on page 100.

The characteristic equation $i_{forw} = f(U_{forw})$ of a p-n junction is now obtained without great effort. First we combine (IV.8.10) with (IV.8.09):

$$i_n(x_p) = \frac{eD_n}{L_n} n_p \left(e^{+\frac{e}{kT}U_{forw}} - 1\right) \qquad \text{(IV.8.11)}$$

In order to arrive at the total forward current i_{forw} the current contribution $i_p(x_p)$ carried by the holes $\oplus$ must be added to (IV.8.11):

$$i_{forw} = i_n(x_p) + i_p(x_p) \qquad \text{(IV.8.12)}$$

This current contribution i_p has essentially the same value at the point $x = x_p$ as at the point $x = x_n$, since the recombination in the transition region is to be neglected:

$$i_p(x_p) \approx i_p(x_n) \qquad \text{(IV.8.13)}$$

In analogy to (IV.8.11) we have

$$i_p(x_n) \approx \frac{eD_p}{L_p} p_n \left(e^{+\frac{e}{kT}U_{forw}} - 1\right) \qquad \text{(IV.8.14)}$$

The combination of (IV.8.12) with (IV.8.13) and (IV.8.14) leads to the characteristic equation for the forward current

$$i_{forw} = e\left(\frac{D_n n_p}{L_n} + \frac{D_p p_n}{L_p}\right)\left(e^{+\frac{e}{kT}U_{forw}} - 1\right)$$

and for the reverse current

$$i_{rev} = e\left(\frac{D_n n_p}{L_n} + \frac{D_p p_n}{L_p}\right)\left(1 - e^{-\frac{e}{kT}U_{rev}}\right) \qquad \text{(IV.8.15)}$$

We obtain, therefore, the same exponential characteristics for the Shockley p-n junction with low recombination as from the diode theory for the metal-semiconductor contact. The saturation current, however, is now

$$i_S = e\left(\frac{D_n n_p}{L_n} + \frac{D_p p_n}{L_p}\right) \qquad \text{(IV.8.16)}$$

With (IV.8.06), the Nernst-Townsend-Einstein relation (IV.4.12), and the mass-action law (IV.6.04) we find

$$i_S = \sqrt{e\,kT} \cdot n_i^2 \cdot \left(\sqrt{\frac{\mu_n}{\tau_n}}\frac{1}{p_p} + \sqrt{\frac{\mu_p}{\tau_p}}\frac{1}{n_n}\right) \qquad \text{(IV.8.17)}$$

which yields, with Eq. (IV.8.05), applicable to the specific reaction

$\ominus + \oplus \rightarrow 0$

$$i_S = \sqrt{e \cdot r \cdot kT}\, n_i^2 \left(\sqrt{\frac{\mu_n}{p_p}} + \sqrt{\frac{\mu_p}{n_n}} \right) \qquad \text{(IV.8.18)}$$

These saturation-current values are much lower than those derived from the diode theory for metal-semiconductor contacts. With the help of (IV.4.07), (IV.8.17), and (VII.9.22) we obtain with $n_B = n_i$ in the diode theory,[1] setting the two terms in (IV.8.17)[2] equal to each other,

$$\frac{i_S \text{ (metal semiconductor)}}{i_S \text{ (p-n junction)}} = \frac{1}{2} \frac{1}{\sqrt{2\pi}} \sqrt{\frac{e\tau_n}{m\mu_n}} \cdot \frac{p_p}{n_i}$$

$$= \sqrt{7.0 \cdot 10^{13} \frac{(\tau_n/\text{sec})}{(\mu_n/\text{cm}^2\,\text{volt}^{-1}\,\text{sec}^{-1})}} \cdot \frac{p_p/\text{cm}^{-3}}{n_i/\text{cm}^{-3}} \qquad \text{(IV.8.19)}$$

For germanium with $\tau_n = 100\ \mu\text{sec}$, $\mu_n = 3{,}500\ \text{cm}^2/\text{volt-sec}$, $p_p = 10^{16}\ \text{cm}^{-3}$, and $n_i = 10^{13}\ \text{cm}^{-3}$, we thus obtain

$$\frac{i_S \text{ (metal semiconductor)}}{i_S \text{ (p-n junction)}} = 1.4 \cdot 10^{+6}$$

We hope that the foregoing discussion has shown that the physical reason for the unipolarity of a p-n rectifier with long diffusion length no longer lies in carrier-concentration extension effects within the transition region. The actual transition region, within which the carrier concentration is approximately equal to the intrinsic density n_i, is no longer of decisive importance for the magnitude of the actual current, for this current is readily provided by small deviations from the Boltzmann equilibrium.[3] The decisive effect is the current yield of the diffusion tails. The current yield of a diffusion tail is, however, vastly different for the two current directions (see Fig. IV.8.3). In

[1] For $n_B < n_i$ an inversion occurs within the boundary region (see p. 26). The neglect of carriers of opposite polarity in the diode theory is no longer permissible in the boundary region.

[2] This is permissible for order-of-magnitude considerations.

[3] For the reverse direction this is, however, approximately correct only to $U_{rev} < V \ln([L/x_0]\sqrt{2/\pi})$, where $V = kT/e = 25.9\ \text{mv}\ (T/300°\text{K})$, $L = L_n$ or L_p is one of the two diffusion lengths, and $x_0 = \sqrt{\varepsilon V/4\pi e n_{D^+}}$ or $\sqrt{\varepsilon V/4\pi e n_{A^-}}$ is the so-called Debye length of the semiconductor. This was calculated by Herlet for the case of constant impurity concentrations n_{A^-} and n_{D^+} changing abruptly at the junction. The term Debye length stems from the similarity of x_0 with the characteristic length in the Debye-Hückel theory of strong electrolytes. As an aid for visualizing this length, we recall that a constant space charge produces a parabolic potential distribution according to the Poisson equation (IV.2.05). At the beginning of this parabolic potential distribution, the potential difference of $\frac{1}{2}V$ is established within a Debye length.

one direction the necessary concentration increases are possible without limit so that currents of any magnitude can be carried. The concentration decrease required for the other current direction, however, rapidly approaches a limit because the concentration at the beginning of the diffusion tail cannot be made less than zero. This should help in visualizing the saturation of the current for reverse polarity.

These p-n rectifiers with long diffusion lengths exhibit excellent reverse characteristics since the diffusion tails with their extremely low concentrations govern the process. This was demonstrated in the comparison with a metal-semiconductor contact. From the practical standpoint, high diffusion lengths are attained by using crystals of greatest possible perfection since recombination takes place primarily at surfaces and crystal imperfections. It is not only necessary to have single crystals, but the single crystals must have a high degree of perfection without dislocations and domain structures.

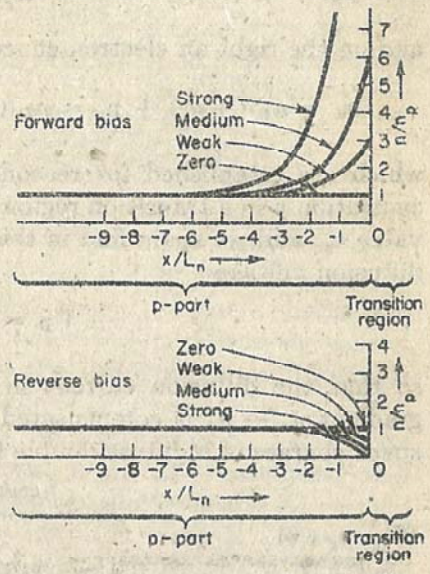

FIG. IV.8.3. Concentration distribution of the electrons in the diffusion tail within the p part. Linear plot.

This leads to an entirely new chapter of the physics of electronic semiconductors which cannot be entered upon here.[1]

§9. Supplementary Remarks on p-n Junctions

a. Abrupt and Gradual Change of Impurity Concentrations in the Junction Region

Up to now we have made the rather unrealistic assumption that the impurity concentrations n_{A^-} and n_{D^+} maintain their constant values up to the center $x = 0$ of the p-n junction, where they drop abruptly to zero. If we examine the reasoning of the previous paragraphs, including §8, we find that this unrealistic assumption does not affect the results appreciably. Assuming, for instance, $n_{A^-}(x)$ and $n_{D^+}(x)$ curves

[1] See in this connection W. Shockley and W. T. Read, Jr., *Phys. Rev.*, **87**: 835 (1952).

as shown in Fig. IV.9.1, we find in analogy to the previously considered case (Fig. IV.6.1) on the left a hole concentration

$$p_p = n_{A^-}(-\infty) + n_p = n_{A^-}(-\infty) + \frac{n_i^2}{p_p} \approx n_{A^-}(-\infty) \quad \text{(IV.9.01)}$$

and on the right an electron concentration

$$n_n = n_{D^+}(+\infty) + p_n = n_{D^+}(+\infty) + \frac{n_i^2}{n_n} \approx n_{D^+}(+\infty) \quad \text{(IV.9.02)}$$

which are established for reasons of neutrality. The electron concentration n in a transition region drops again from the value n_n to the value n_p, and we again find in this transition region for zero current a diffusion voltage

$$V_D = \frac{kT}{e} \ln \frac{n_n}{n_p} \quad \text{(IV.9.03)}$$

so that the diffusion current in the direction of the concentration gradient $n_n \rightarrow n_p$ is compensated by an opposite field current. The space charges of a diffuse double layer are necessary for the formation

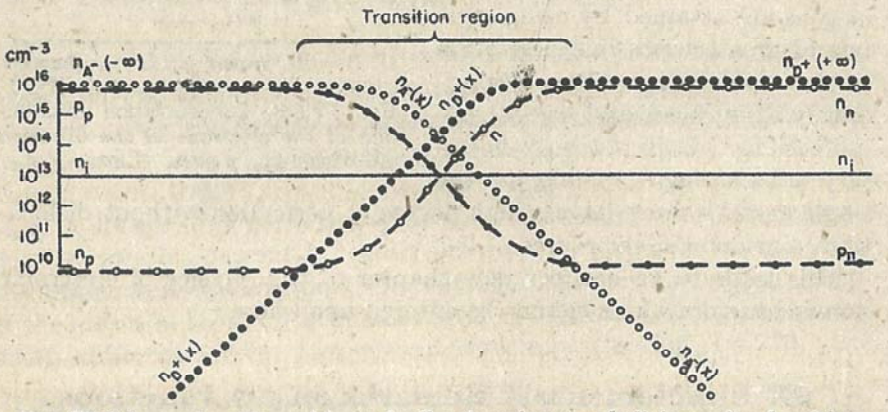

FIG. IV.9.1. p-n junction with gradually changing (graded) impurity densities.

of the potential step V_D. Hence, it is not possible for n and p to have, in the entire transition region, values prescribed by the neutrality condition

$$n + n_{A^-} = p + n_{D^+} \quad \text{(IV.9.04)}$$

and the law of mass action

$$n \cdot p = n_i^2 \quad \text{(IV.9.05)}$$

These would be

$$n_{\text{neutral}}(x) = +\tfrac{1}{2}[n_{D^+}(x) - n_{A^-}(x)] + \sqrt{\tfrac{1}{4}[n_{D^+}(x) - n_{A^-}(x)]^2 + n_i^2} \quad \text{(IV.9.06)}$$

$$p_{\text{neutral}}(x) = -\tfrac{1}{2}[n_{D^+}(x) - n_{A^-}(x)] + \sqrt{\tfrac{1}{4}[n_{D^+}(x) - n_{A^-}(x)]^2 + n_i^2} \quad \text{(IV.9.07)}$$

For *abrupt* transitions of the impurity concentrations n_{A^-} and n_{D^+} (extreme case: Fig. IV.6.1), the deviations $p(x) - p_{\text{neutral}}(x)$ and $n(x) - n_{\text{neutral}}(x)$ will have the order of magnitude of $n_{A^-}(x)$ or $n_{D^+}(x)$, respectively. For very gradual transitions, these deviations $p(x) - p_{\text{neutral}}(x)$ and $n(x) - n_{\text{neutral}}(x)$ will be small compared to $n_{A^-}(x)$ or $n_{D^+}(x)$, respectively. In both cases, however, the same diffusion voltage V_D prescribed by (IV.9.03) must be formed in a double layer—in one case in a relatively thin layer with space-charge densities $\rho(x) \approx -en_{A^-}(x)$ or $\rho(x) \approx +en_{D^+}(x)$, in the other case in a layer with the space-charge densities $\rho(x) \ll -en_{A^-}(x)$ or $\rho(x) \ll +en_{D^+}(x)$.

These details do not, at least at this stage, prevent a distinction between the transition region and the two diffusion tails in the case of current flow. For current flow, too, the Boltzmann equilibrium is approximately maintained. In the regions adjacent to the transition region, the dominant carrier type (the "majority carriers") carries its current contribution as pure field current in a small "base field" and the minority carrier type (the "minority carriers") carries its current contribution in the form of a pure diffusion current. Therefore, Eq. (IV.8.09) applies again for the current contribution of the electrons beyond the transition region. The concentration value at the beginning of the diffusion tail is again determined by the Boltzmann distribution in the adjacent transition region [Eq. (IV.8.10)]. Thus we find again (IV.8.11) for the current contribution i_n of the electrons and finally (IV.8.15) for the total current.

The realization that the validity of the characteristic equation (IV.8.15) depends solely on the adjacency of a Boltzmann region and two diffusion tails, and not on the variation of the concentration within the Boltzmann region, will be fruitful when we consider unsymmetrical p-n junctions later in this paragraph. However, we should point out here that the maximum field strength is of course appreciably higher in a thin transition region than in a wide one, since the height of the potential step is always given, in the critical reverse case, by $V_D + U_{\text{rev}}$ independent of the width. Experience indicates that at field strengths of 10^5 to 10^6 volt cm^{-1} breakdown occurs, which may lead to destruction of the crystal. However, reversible secondary phenomena precede this stage; we shall mention here only the so-called Zener effect,[1] where valence electrons are transferred into the conduction band by high field strengths. K. B. McAfee, E. J. Ryder,

[1] See, for instance, pp. 223ff.

W. Shockley, and M. Sparks[1] have explained the reversible steep rise of the reverse current above a critical reverse voltage by the Zener effect.[2] Disregarding the actual mechanism, it is known that the failure of rectifiers above a certain reverse voltage is associated with excessive field strengths, and this shows that the impurity transition must be as gradual as possible in order to attain the highest possible reverse voltages.[3] The only limitation in this direction is the original assumption that the transition region must be small compared with the diffusion length which, in turn, calls for the greatest possible diffusion lengths. This requires the use of single crystals with very low imperfection content, as pointed out previously.

b. Asymmetrical p-n junctions

We have already pointed out that the characteristic Eqs. (IV.8.15) and (IV.8.17) are based on the adjacency of a Boltzmann region and two diffusion tails and that the details of the concentration distributions, particularly within the Boltzmann region, are unimportant. This allows us to apply the characteristic Eqs. (IV.8.15) and (IV.8.17) to the unsymmetrical p-n junction.

We dope, for instance, the p region with 10^{18} acceptors/cm^3 and the n region with only 10^{15} donors/cm^3. Then we have

$$p_p = 10^{18}\ \text{cm}^{-3} \gg n_n = 10^{15}\ \text{cm}^{-3}$$

The first term of the sum in the expressions (IV.8.16) and (IV.8.17) for the saturation current represents the electrons and disappears for all practical purposes in comparison with the hole contribution. Since the saturation current is not only decisive in the reverse direction but determines the *entire* characteristic (IV.8.15), the forward current consists essentially of holes which flow from the p germanium to the n germanium. The forward-biased p-n junction of a highly doped p

[1] K. B. McAfee, E. J. Ryder, W. Shockley, and M. Sparks. *Phys. Rev.*, **83**: 650 (1951). The exponent in Eq. (1) of this paper is, however, too large by a factor 2. This appears to be only a printing error, for in the numerical equation (3) of the same paper the exponent has again the right value.

[2] See also G. K. McKay and K. B. McAfee, *Phys. Rev.*, **91**: 1079 (1953).

[3] R. N. Hall and W. C. Dunlap, *Phys. Rev.*, **80**: 467 (1950). We would like to suggest the use of the expression *p-n junction* also for very steep impurity concentration variations, i.e., for very abrupt transitions between p and n germanium. As a criterion for the distinction between *transition* or *junction* and *contact* or *boundary*, we shall use the uniformity or change of the crystal lattice. W. Shockley had probably the same distinction in mind when in his original paper [*Bell System Tech. J.*, **28**: 444 (1949)], he considered abruptly changing impurity distributions under the heading "junction."

and lightly doped n material acts, therefore, as a good "emitter" of holes into the n material. This will be of importance in transistor physics.

In this connection we mention another fact which is not restricted to the asymmetry of the p-n contact but is also important for the mechanism of a p-n-p transistor. The reverse current of a p-n junction is small because of the small current yield of the diffusion tails with reverse bias. If, therefore, in a p-n junction with reverse bias, carrier pairs are introduced by other means than the normal *thermal* excitation such as light incidence (internal photoelectric effect) or by high field strengths (Zener effect), the reverse current will be affected by these additional minority carriers. The same will apply if the carrier depletion is alleviated by the injection of carriers from a foreign carrier source. This means that a p-n junction with a reverse bias acts as a good "collector." We shall treat these matters in more detail in the next chapter.

§10. Junction Capacitance

It is clear from the preceding paragraphs that for any given voltage there is a certain distribution of charge carriers at and near the p-n junction. If the voltage is changed, this charge distribution has to be readjusted, too. This means that, during a change in voltage, the current is not equal to the current which corresponds to the instantaneous voltage but that there is an additional current which is proportional to the *rate* of change of voltage. Such a current is a capacitive current, and we therefore have to attribute a certain capacitance to a p-n junction.

The mobile charges in a semiconductor are minority carriers and majority carriers. Both contribute to the junction capacitance, but in a markedly different way. We treat the minority carriers first; and we restrict ourselves to junctions with low internal recombination (Shockley case).

When a p-n junction is biased in the forward direction, a diffusion tail of minority carriers extends to a depth of about a diffusion length into the bulk semiconductor, as shown for the p-type side in Fig. IV.8.3.

On the p side the electron density of the diffusion tail—that is, the total electron density minus the equilibrium density—is, according to Eqs. (IV.8.07) and (IV.8.10),

$$n(x) - n_p = n_p \left(e^{\frac{e}{kT}U_{\text{forw}}} - 1\right) e^{-\frac{x-x_p}{L_n}} \qquad \text{(IV.10.01)}$$

The total number of electrons stored per unit area in this diffusion tail follows by integration

$$\frac{Q_n}{e} = \int_{-\infty}^{x_p} [n(x) - n_p]\, dx = n_p L_n \left(e^{\frac{e}{kT} U_{forw}} - 1\right) \quad \text{(IV.10.02)}$$

If now the voltage is changed by dU_{forw}, the stored electron charge changes by dQ_n, resulting in what may be termed the *differential storage capacitance for electrons*

$$C_n = \frac{dQ_n}{dU_{forw}} = \frac{e^2 n_p L_n}{kT} e^{\frac{e}{kT} U_{forw}} \quad \text{(IV.10.03)}$$

In addition, we have a similar differential storage capacitance for holes

$$C_p = \frac{dQ_p}{dU_{forw}} = \frac{e^2 p_n L_p}{kT} e^{\frac{e}{kT} U_{forw}} \quad \text{(IV.10.04)}$$

so that the total storage capacitance is

$$C_{stor} = C_n + C_p = \frac{e^2}{kT} (n_p L_n + p_n L_p)\, e^{\frac{e}{kT} U_{forw}} \quad \text{(IV.10.05)}$$

We have called this capacitance a differential capacitance because it is not the total charge divided by the voltage but rather the change in the charge for a small change in the voltage. Whenever, as in our case, the charge is not directly proportional to the voltage, this is a more useful definition because it is the differential capacitance which is seen by a small a-c signal superimposed on a d-c signal.

The storage capacitance is also often called diffusion capacitance because it arises from the carriers diffusing across the junction.

From (IV.10.05) it follows that the storage capacitance increases rapidly with the forward bias. The forward current increases in a similar fashion, and for $eU_{forw} \gg kT$ the storage capacitance is directly proportional to the current.

For a reverse bias, the storage capacitance drops off very rapidly to zero. However, the total capacitance of the junction does not vanish; there is an additional capacitance of the space-charge transition region itself, generally called the transition capacitance. This is the capacitance due to the *majority* carriers which was mentioned earlier. As we will show, it also varies with voltage, but much less so; in the forward direction, the transition capacitance is normally much smaller than the storage capacitance so that it can be neglected (see the problems in §11). We therefore can restrict our discussion of the transition capacitance to the reverse direction, where the opposite is true.

As shown already in Fig. IV.3.4 for the metal-semiconductor contact,

the space-charge region widens for increasing reverse bias. The same is true, of course, for a p-n junction. This widening arises from the displacement of the majority carriers which neutralize the ionized impurities. As a result, a capacitive current flows. But contrary to the case of the minority carrier storage capacitance, the capacitive current of the transition capacitance is not a *carrier* current across the junction but rather a *displacement* current as in an ordinary condenser.

This similarity between the transition capacitance and an ordinary condenser capacitance holds quantitatively. To show this, we introduce the concept of an "equivalent condenser" that corresponds to a reverse-biased p-n junction. This is an ordinary plate condenser which is filled with a dielectric of the same dielectric constant as the semiconductor and which has a plate distance equal to the width of the space-charge region of the junction. If the boundaries of the space-charge layer are parallel planes, the equivalent condenser is a parallel-plate condenser; if they are curved in any arbitrary way, the plates of the equivalent condenser may be assumed to be of the same shape. The equivalent condensers that correspond to different reverse-bias values of the same junction are, of course, different.

If we now apply a small bias dU_{rev} across the equivalent condenser, its plates will charge up by a small amount dQ. Let us then take away this charge from the plates of the condenser and transfer every charge element, point by point, to the equivalent position at the space-charge layer boundary of the junction. This can be done by displacing this boundary by an infinitesimal amount. These charge elements are then arranged in the semiconductor in an identical geometry and in a medium of identical dielectric constant as before in the condenser. They therefore produce—superposed on the already present field distribution inside the space-charge layer—an additional field that is identical with the field inside the condenser, because the original charge distribution itself is not affected by the new charges. The additional voltage that develops across the p-n junction is then equal to the dU_{rev} that had been applied across the condenser. This means that the differential transition capacitance dQ/dU_{rev} is equal to the capacitance of the equivalent plate condenser.

For parallel-junction boundaries with a plate separation $x_n - x_p$, the capacitance per unit area is then

$$C_{tr} = \frac{\varepsilon}{4\pi(x_n - x_p)} \qquad \text{(IV.10.06)}$$

The problem therefore reduces to a determination of the space-charge region width as a function of the voltage.

This problem can be solved as follows: The space-charge region of a p-n junction is an electric double layer containing an equal number of positive and negative charges. Therefore, if ρ is the local space-charge density

$$\int_{x_p}^{x_n} \rho \, dx = 0 \qquad \text{(IV.10.07)}$$

The electrical dipole moment per unit area of the layer is

$$M = \int_{x_p}^{x_n} \rho x \, dx$$

and the potential difference ΔV across such a double layer is known to be equal to

$$\Delta V = \frac{4\pi}{\varepsilon} M$$

In a p-n junction biased in the reverse direction, the total potential difference is $U_{rev} + V_D$, where V_D is the diffusion potential. Therefore,

$$U_{rev} + V_D = \frac{4\pi}{\varepsilon} \int_{x_p}^{x_n} \rho x \, dx \qquad \text{(IV.10.08)}$$

If $\rho(x)$ were known, one could determine x_n and x_p and thereby $x_n - x_p$ from the two equations (IV.10.07) and (IV.10.08).

From the interior of a reverse-biased junction, all mobile carriers are swept out. The charge density, then, is equal to the charge density of the ionized impurities and, if all impurities are ionized,

$$\rho(x) = e[n_D(x) - n_A(x)] \qquad \text{(IV.10.09)}$$

Near the ends of the boundaries of the space-charge layer, however, the mobile carriers are only partially swept out and the charge density approaches zero gradually, as shown in Fig. IV.3.4. This tailing off takes place in that region where the electrostatic potential differs from the potential of the adjoining neutral semiconductor by only a few kT/e. For every deviation kT/e from the outside potential, the majority carrier density drops by e^{-1} so that the remaining mobile carrier density is negligible as soon as the potential has changed by more than a few kT/e. Now, kT/e is a very small voltage, at room temperature about 0.026 volt. This tail region, therefore, contributes only a small amount of the total potential difference whenever $U_{rev} + U_D$ is large compared to kT/e, and this is practically always the case, since generally V_D alone is large compared to kT/e.

It has therefore been suggested by Schottky,[1] and justified by a detailed mathematical analysis, that one can replace the actual gradual drop-off in the space charge with negligible error by an abrupt drop-off.[2] This means that x_p and x_n are determined from the two equations

$$\int_{x_p}^{x_n} [n_D(x) - n_A(x)]\, dx = 0 \tag{IV.10.10}$$

$$\int_{x_p}^{x_n} [n_D(x) - n_A(x)] \cdot x\, dx = \frac{\varepsilon}{4\pi e} (U_{\text{rev}} + V_D) \tag{IV.10.11}$$

Actually, it is this introduction of an abrupt transition that gives a well-defined meaning to the quantities x_p and x_n; with a gradual transition, it is not immediately clear which coordinate is to be chosen as the "end" of the space-charge layer. Schottky showed that the values derived from the last two equations actually are the proper values.

Equations (IV.10.10) and (IV.10.11) can be solved in closed form only for a number of simple distributions. In the case of an abrupt transition, say

$$n_D(x) - n_A(x) = \begin{cases} -n_A \\ +n_D \end{cases} \quad \text{for} \quad \begin{cases} x < 0 \\ x > 0 \end{cases} \tag{IV.10.12}$$

we find

$$-x_p = \sqrt{\frac{\varepsilon}{2\pi e} \frac{n_D}{n_A (n_D + n_A)} (U_{\text{rev}} + V_D)} \tag{IV.10.13}$$

$$x_n = \sqrt{\frac{\varepsilon}{2\pi e} \frac{n_A}{n_D (n_D + n_A)} (U_{\text{rev}} + V_D)} \tag{IV.10.14}$$

$$x_n - x_p = \sqrt{\frac{\varepsilon}{2\pi e} \frac{n_D + n_A}{n_D n_A} (U_{\text{rev}} + V_D)} \tag{IV.10.15}$$

An important special case of this is the extremely unsymmetrically doped junction because it occurs in many practical alloy junctions and describes the behavior of the semiconductor side of a metal-semiconductor contact as well. For $n_A \gg n_D$, we obtain

$$-x_p = \sqrt{\frac{\varepsilon}{2\pi e} \frac{n_D}{n_A^2} (U_{\text{rev}} + V_D)} \tag{IV.10.16}$$

$$x_n = \sqrt{\frac{\varepsilon}{2\pi e} \frac{1}{n_D} (U_{\text{rev}} + V_D)} \qquad \text{independent of } n_A \tag{IV.10.17}$$

$$x_n - x_p \approx x_n \tag{IV.10.18}$$

[1] W. Schottky, *Z. Physik*, **118**: 539 (1942).

[2] Note that Fig. IV.3.4 has a logarithmic scale. If replotted linearly, it shows that the real charge distribution drops off rather abruptly.

Since the junction width always increases with increasing voltage, it follows already from (IV.10.06) that the transition capacitance decreases with increasing U_{rev}. But while the storage capacitance decreases exponentially, the decrease of C_{tr} is fairly slow. In our example we found a decrease with the inverse square root of the total potential difference. Other impurity distributions lead to other variations. For further details we refer the reader to the problems in §11.

§11. Problems

1. In analogy to a differential capacitance dQ/dU, a differential conductance can be defined as di/dU. Give a formula for the differential conductance of a p-n junction both as a function of the voltage and as a function of the current.

2. An abrupt p-n junction in germanium may have the following physical parameters:

$$n_A = 10^{16}\ \text{cm}^{-3} \qquad n_D = 10^{15}\ \text{cm}^{-3} \qquad \tau_n = 10\ \mu\text{sec} \qquad \tau_p = 50\ \mu\text{sec}$$

Assuming the values for n_i, ε, μ_n, and μ_p given in the text, calculate the following electrical properties of this junction:

a. The diffusion voltage V_D.

b. The thickness of the space-charge region and the transition capacitance per unit area for zero bias.

c. The diffusion capacitance per unit area for zero bias.

d. The reverse saturation current density and the forward current density for a bias of 0.2 volt.

3. For the p-n junction in the preceding problem, determine graphically or numerically the voltage for which the diffusion capacitance becomes equal to the transition capacitance.

4.* The following experimental data are known about a particular germanium p-n junction:

a. The transition capacitance per unit area follows the relationship:

$$\frac{1}{C^2} = 1.77 \cdot 10^{-8} \left(\frac{U_{rev}}{1\ \text{volt}} + 0.314 \right) \frac{\text{cm}^4}{(\mu\text{f})^2}$$

b. The reverse saturation current at room temperature is 192 μa/cm^2.

c. The diffusion capacitance per unit area at zero bias is 0.363 μf/cm^2.

Determine from these values the physical parameters of the junction, such as the impurity densities and the lifetimes, assuming the junction to be abrupt and assuming the values for n_i, ε, μ_n, and μ_p given in the text.

What is the uncertainty in the physical parameters if the contact potential varies by ± 0.001 volt, provided that the other experimental data are accurate?

5. Verify Eqs. (IV.10.13) to (IV.10.15).

6. What is the capacitance-voltage relationship for a graded junction for which the impurity density is given by

$$n_D(x) - n_A(x) = ax \qquad \text{(IV.11.01)}$$

where a is a constant of the dimension cm^{-4}? Assume, for example, a value of $a = 10^{19}$ cm^{-4}. At what voltage, then, is the capacitance equal to 10^4 $\mu\mu$f/cm^2?

7. What is the capacitance-voltage relationship for a so-called p-i-n junction,

that is, a junction where the p region and the n region are separated by an intrinsic region of a certain thickness, say w_i?

$$n_D(x) - n_A(x) = \begin{cases} -n_A & x < 0 \\ 0 & \text{for } 0 < x < w_i \\ +n_D & w_i < x \end{cases} \qquad \text{(IV.11.02)}$$

How does the result differ from the behavior of an abrupt junction without the intrinsic layer but with otherwise identical structure? Calculate the zero bias capacitance for a germanium junction with $n_D = 10^{15}$ cm^{-3}, $n_A = 10^{18}$ cm^{-3}, $w_i = 10^{-4}$ cm.

CHAPTER V

The Physical Mechanism of Crystal Amplifiers (Transistors)

§1. Introduction

The transistor is a device for the amplification of electrical signals; so it may not be too far-fetched to refer in the discussion of its mechanism to the oldest and simplest device of this type, namely, the electromagnetic telegraph relay (see Fig. V.1.1). Here, a weak current arriving over a long transmission line actuates a switch which starts or blocks the strong current from a local source. The essential part of this process can be described as a signal which varies the conductivity in a current path of a local current source, effected in this particular case by a variation of the cross section at a certain point.

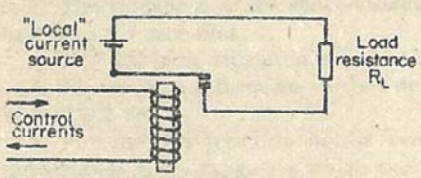

FIG. V.1.1. Amplification by a telegraph relay.

As an alternative of such a modification of the geometric dimensions of the current path, one could also consider a change of the specific conductivity, for instance, by varying the number of carriers. This actually happens in a number of transistor types, namely, by more or less intensive injection of additional charge carriers. Within the range of this common mechanism, the various types differ only in the nature of the affected current path. If the injection occurs into an ohmic conductor, we have the filamentary transistor (§2). If the affected current path is a *p-n* junction, we have the *n-p-n* transistor (§3). Finally, in the point-contact transistor, the boundary layer of a metallic point contact is affected by injected carriers (§4). The last transistor type to be discussed, namely, the unipolar transistor, operates like the relay by means of continuously variable geometrical dimensions of the current path (§5). This can again be interpreted as a variation of the carrier concentration as in the other transistor types. In contrast to the other types, it is here restricted to more or less

extended boundary regions of the current cross section. Furthermore, this variation applies to the majority carriers and not the minority carriers.

§2. The Filamentary Transistor

As was pointed out in the introduction, in this transistor type additional charge carriers are injected into an ohmic conductor. So far, we have used the term injection as though it required no further explanation. The error of this notion is evident from the fact that effective carrier injection is possible only into a semiconductor, but not into a metal. Therefore we must first discuss carrier injection and the associated time effects.

a. Time Effects with Carrier Injection

We choose the example of a grounded germanium crystal which is doped with $n_{D^+} = 10^{16}$ arsenic atoms/cm^3 and is, therefore, an n-type semiconductor.[1] The neutrality condition requires in this case an electron concentration n of

$$n = n_{D^+} \approx 10^{+16}\ \mathrm{cm}^{-3} \tag{V.2.01}$$

Now we want to increase the concentration n by electron bombardment, for instance, by $\delta n = 10^{15}$ cm^{-3} to $n + \delta n = 1.1 \cdot 10^{16}$ cm^{-3}. This can be accomplished for only an extremely short period of time because the additionally introduced electrons repel each other and flow to ground. Expressing this more precisely, the introduction of $\delta n = 10^{+15}$ electrons/cm^3 disturbs the neutrality of the conductor, and the resulting space charge ρ creates an electric field which sets the entire electron concentration $n + \delta n$ in motion. The resulting currents reduce the additional concentration again to zero. Not only the few additional δn electrons participate in this process but predominantly the $n(\gg \delta n)$ electrons which were already present before the disturbance. This disturbance is, therefore, rapidly removed and can exist only for a very short time.

In a quantitative treatment of the time decay of a space charge $\rho(t)$ we start from the continuity equation:

$$\frac{\partial \rho}{\partial t} = -\ \mathrm{div}\ \mathbf{i} \tag{V.2.02}$$

[1] Substitutional imperfections or impurities in germanium can be assumed to be fully ionized—except at extremely low temperatures and very high doping concentrations. Hence we can put $n_D = n_{D^+}$.

With the current

$$\mathbf{i} = \sigma \mathbf{E} = -\sigma \operatorname{grad} V \tag{V.2.03}$$

we obtain from (V.2.02)[1]

$$\frac{\partial \rho}{\partial t} = +\sigma \cdot \Delta V \tag{V.2.04}$$

Combination with Poisson's equation

$$\Delta V = -\frac{4\pi}{\varepsilon}\rho \tag{V.2.05}$$

results in

$$\frac{\partial \rho}{\partial t} = -\frac{4\pi\sigma}{\varepsilon}\rho \tag{V.2.06}$$

with the solution

$$\rho(t) = \rho(0)\, e^{-\frac{t}{T_{relax}}} \tag{V.2.07}$$

from which we find the relaxation time

$$T_{relax} = \frac{\varepsilon}{4\pi\sigma} \tag{V.2.08}$$

The germanium with 10^{16} cm^{-3} electrons just mentioned would have a conductivity

$$\begin{aligned}\sigma &= e\mu_n n \\ &= 1.6 \cdot 10^{-19} \text{ coulomb} \cdot 3.6 \cdot 10^{+3} \text{ cm}^2/\text{volt-sec} \cdot 10^{+16} \text{ cm}^{-3} \\ &\qquad = 5.76 \text{ ohm}^{-1} \text{ cm}^{-1} \\ &= 5.76 \cdot 9 \cdot 10^{+11} \text{ sec}^{-1} = 5.19 \cdot 10^{+12} \text{ sec}^{-1}\end{aligned} \tag{V.2.09}$$

With $\varepsilon_{Ge} = 16.2$ we find from (V.2.08) in this case

$$T_{relax} = 2.49 \cdot 10^{-13} \text{ sec} \tag{V.2.10}$$

From (V.2.08) and (V.2.09) we obtain the relation for the relaxation time

$$T_{relax} = \frac{\varepsilon}{4\pi e \mu_n n} \tag{V.2.11}$$

which shows that the speed of the decay process is determined not by the disturbance itself but by the undisturbed concentration n. T_{relax} becomes very small because of the magnitude of n. Hence, if one introduces electrons into an n-type conductor, the change of concentration is maintained for only 10^{-12} to 10^{-13} sec.

[1] With $\Delta = \operatorname{div} \operatorname{grad} = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$.

The situation is, however, vastly different if $\delta p = 10^{15}$ holes/cm³ can be introduced additionally into the n germanium under consideration. These additional holes also repel each other, but only their own small concentration $\delta p = 10^{15}$ cm^{-3} is available for the decay of the concentration increase δp. Therefore the slow decay process does not take place at all. Instead, the electrons increase their concentration $n = 10^{16}$ cm^{-3} by $\delta n = 10^{15}$ cm^{-3} to $1.1 \cdot 10^{16}$ cm^{-3}; because their own high concentration 10^{16} cm^{-3} is available for this build-up process it proceeds very rapidly. Neutrality is again established as soon as the electrons have increased their concentration $1.0 \cdot 10^{16}$ cm^{-3} to $1.1 \cdot 10^{16}$ cm^{-3}, and there are no fields left which could remove any electrons or holes.

However, this condition is not maintained indefinitely either. In Chap. I, §3, we have shown that in all semiconductors a thermally determined carrier generation g and a carrier annihilation $r \cdot n \cdot p$ by recombination counteract each other continually. Time changes in carrier concentrations $n(t)$ and $p(t)$ must follow the law[1]

$$\frac{dn(t)}{dt} = \frac{dp(t)}{dt} = g - r \cdot n(t) \cdot p(t) = r[n_i^2 - n(t) \cdot p(t)] \quad \text{(V.2.12)}$$

The intrinsic density n_i in this law has approximately[2] the value 10^{13} cm^{-3} in germanium at room temperature. In equilibrium a hole concentration $p = 10^{10}$ cm^{-3} must coexist[3] with the electron concentration $n = 10^{16}$ cm^{-3}, because the time-independent equilibrium concentrations n and p are, according to (V.2.12),

$$0 = g - r \cdot n \cdot p = r(n_i^2 - n \cdot p) \quad \text{(V.2.13)}$$

The time-dependent disturbances $\delta n(t) = \delta p(t) \approx 10^{11}$ cm^{-3} introduced into (V.2.12) lead, in conjunction with (V.2.13), to

$$\begin{aligned} \frac{d}{dt}\delta n = \frac{d}{dt}\delta p &= r[n_i^2 - (n + \delta n)(p + \delta p)] \\ &= -r(n\,\delta p + p\,\delta n + \delta n\,\delta p) \end{aligned}$$

In view of the relative orders of magnitude of $n = 10^{16}$ cm^{-3},

[1] Provided concentration changes are not effected by other causes such as the divergence of a carrier flow.

[2] E. M. Conwell gives $n_i = 2.5 \cdot 10^{13}$ cm^{-3}. *Proc. IRE*, **40**: 1329 (1952), Table II.

[3] The concentration $p = 10^{10}$ cm^{-3} does not play any role in the neutrality condition compared with $n = 10^{16}$ cm^{-3} and $n_{D^+} = 10^{16}$ cm^{-3}. Even in the problem of the dispersion of the injected holes $\delta p = 10^{15}$ cm^{-3}, it is for all practical purposes not necessary to consider the already present equilibrium density $p = 10^{10}$ cm^{-3}.

$p = 10^{10}$ cm^{-3}, and $\delta n = \delta p \approx 10^{11}$ cm^{-3}, we obtain

$$\frac{d}{dt}\delta n = \frac{d}{dt}\delta p = -rn\,\delta p = -\frac{\delta n}{\tau_p} \qquad \text{(V.2.14)}$$

$$\delta n = \delta p \sim e^{-\frac{t}{\tau_p}} \qquad \text{(V.2.15)}$$

We can now see that even a *neutral* deviation $\delta n = \delta p$ from thermal equilibrium ($n = 10^{16}$ cm^{-3}, $p = 10^{10}$ cm^{-3}) cannot last forever but decays exponentially with a "lifetime $\tau_p = 1/(rn)$ of the holes in the n conductor." This lifetime τ_p and also the lifetime $\tau_n = 1/(rp)$ of electrons in a p conductor are strongly dependent on the perfection of the crystal lattice through the recombination coefficient[1] r. Exceptionally perfect crystals exhibit lifetimes as high as 10^{-3} sec, and even relatively poor single crystals rarely have lifetimes less than 10^{-7} sec. The lifetimes τ_p and τ_n are therefore much higher than the relaxation times T_{relax}.

Summarizing and generalizing we can now state: The electrons in n semiconductors and the holes in p semiconductors, i.e., the "majority carriers," remove disturbances of quasi-neutrality in a semiconductor within extremely short times $T_{relax} \approx 10^{-13}$ sec. It is immaterial here how the particular disturbance has come about. If the disturbance is, for instance, caused by the injection of "minority carriers," the majority-carrier concentration is increased within the short time T_{relax} in order to establish neutrality. Both concentrations decay thereafter exponentially with the lifetime τ_{minor} as the time constant, which is very large compared with the relaxation time T_{relax}.

b. The Filamentary Transistor

The foregoing discussion shows that it is useless to inject majority carriers in order to influence the conductivity of a current path. The additional carrier concentrations decay within much too short times T_{relax} or within much too short distances $v_{drift} \cdot T_{relax}$ if the injected carriers are transported with a drift velocity v_{drift}.

When minority carriers are injected, however, the equalizing flow of majority carriers reestablishes neutrality within a few relaxation times T_{relax} so that the space-charge field with its dissipating tendency is removed. The conductance in the current path in question is thus increased by the additional minority carriers as well as by the neutralizing concentration rise of the majority carriers.

The question arises now as to how the injection of minority carriers is accomplished. The introduction by bombardment from the outside

[1] See also footnote 1, p. 99.

requires a vacuum, high voltages, and electron-optical structures and is, therefore, very cumbersome. Furthermore, this method would be restricted to electrons and could be applied only to a p-type semiconductor. The following two methods are much more elegant. First, light excitation may be used to increase the pair generation above its thermally determined value. The creation of *pairs* of positive and negative carriers leads not even to a momentary deviation from neutrality. This effect is utilized in the so-called phototransistor. Second, charge displacement effects known from rectifier theory may be utilized. Thus holes may be carried from a p semiconductor into an n semiconductor. This is the emitter action of a p-n junction biased in the forward direction, as mentioned on page 106. Finally, it has been

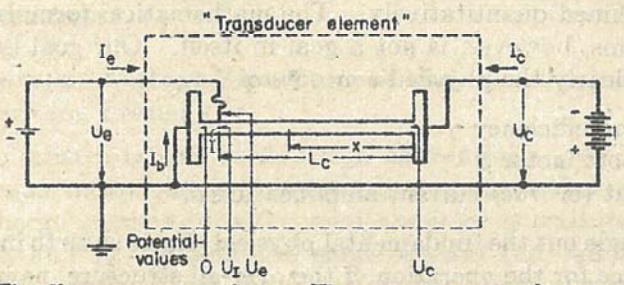

FIG. V.2.1. The filamentary transistor. The current arrows do not represent the direction of the actual currents but rather the direction in which the current in question is considered positive. Accordingly, the voltage arrows do not represent the actual potential drop but rather the direction in which the potential drop (voltage) in question is considered positive. The schematic batteries, however, are represented with the correct polarities as they are applied in the operation of the transistor.

found empirically that the forward current of a metal-semiconductor contact on germanium may consist largely—possibly entirely—of minority carriers. The interpretation of this effect is believed to lie in a disguised p-n action.

Be that as it may, forward-biased p-n junctions and metal-semiconductor point contacts represent simple "emitters" for the injection of minority carriers.

In conjunction with the introductory remarks of §1 we obtain the following arrangement of a crystal amplifier (see Fig. V.2.1). A rod or filament of single-crystal n germanium is provided with *large-area* electrodes at each end to make good ohmic connections. The left "base electrode" is grounded, and a negative bias $[|U_c| \gg + kT/e]$ is applied to the right "collector electrode." An emitter is applied to the rod near the base with a positive bias with respect to the base in order to produce emitter action. A fraction γ of the current I_e which

flows from the emitter into the germanium rod consists of holes (γ = injection efficiency = the fraction of the emitter current carried by minority carriers). These holes are collected by the negative collector after entering the n germanium so that they modulate the conductivity of the current path between emitter and collector, depending on their number. Therefore, variation of the emitter voltage U_e must be capable of controlling the power supplied by the battery in the collector circuit.

c. Survey of Procedure and Objective of the Following Calculations

The preceding rough description of the effects under discussion will now be refined quantitatively. The mathematical formulation of the mechanisms, however, is not a goal in itself. Our goal is, rather, to develop clearly the physical concepts of

Injection efficiency γ
Transport factor β
Inherent (or *true*) current amplification α_i

and to single out the fundamental physical effects so as to indicate their significance for the operation of the over-all structure, namely, for the amplification properties of the transistor.

The transistor is a special case of a switching or rather a transducer element in which a primary or input circuit controls a secondary or output circuit. The behavior of such a transducer element is, accordingly, determined by two current-voltage relations (namely, one for the primary and one for the secondary circuit):

$$U_e = f(I_e, I_c) \tag{V.2.16}$$
$$U_c = g(I_c, I_e) \tag{V.2.17}$$

For small deviations u_e, u_c, i_e, i_c, from an operating point, these equations can be linearized:

$$u_e = r_{11} i_e + r_{12} i_c \tag{V.2.18}$$
$$u_c = r_{21} i_e + r_{22} i_c \tag{V.2.19}$$

where r_{11} and r_{22} are the differential primary and secondary resistances. The coupling resistance r_{21} furnishes the desired influence of the primary circuit on the secondary circuit. The feedback resistance r_{12} determines the feedback effect of the secondary circuit on the primary circuit.

The amplification and transfer properties, respectively, of a circuit element with a primary and a secondary circuit are described appropri-

ately by the short-circuit current amplification, the open-circuit voltage amplification, and the power amplification in the matched condition. These amplification values are generally derived from the resistances $r_{11} \cdots r_{22}$. This will be done in §6, since the relations in question and their derivation have nothing to do with the physical mechanism of the filamentary transistor.

The following calculations for the filamentary transistor relate the values of the resistances $r_{11} \cdots r_{22}$ to the structural data and the characteristic constants of this type of transistor. The equations of §6 are used in the same way for the amplification factors. A discussion of the equations which we derive will help us to understand the importance of the individual effects for the performance of the filamentary transistor.

d. The Current-Voltage Equations of the Filamentary Transistor

First we must establish certain sign conventions. We adopt the choice of signs customary in transistor literature, so that the flow of positive charge carriers into the semiconductor constitutes positive currents I_e and I_c. Within the single crystal rod, we measure a coordinate x positive from the collector end $x = 0$ toward the left (see Fig. V.2.1). The electrical field strength E within the crystal is taken as positive in the direction of increasing x values as usual.

We shall now consider a certain location x of the single crystal rod between the point I (immediately in front of the emitter) and the collector c (see Fig. V.2.1). Here we find a field strength directed from the left to the right, and hence negative according to our convention. Its value is $E_0 + E_1(x) < 0$; E_0 is the component present before injection. The corresponding electron concentration is $n_0 + n_1$ and the hole concentration $n_i^2/n_0 + p_1(x) \approx p_1(x)$, because the injected hole concentration $p_1(x)$ even at low injection levels is large compared with the already present hole density n_i^2/n_0 ($= 10^{10}$ cm^{-3} in our example). Restricting the analysis to low injection levels where we can neglect terms of the second order, we find for the current density I_c/Q of the collector current (Q being the cross-sectional area of the germanium rod):

$$\frac{1}{Q} I_c = e\mu_n n_0 E_0 + e\mu_n n_1(x) E_0 + e\mu_p p_1(x) E_0 + e\mu_n n_0 E_1(x) \quad \text{(V.2.20)}^1$$

[1] The signs in this equation are correct, since I_c and $E_0 + E_1$ are positive in the same direction, namely, from right to left. In transistor operation I_c as well as $E_0 + E_1$ are negative, at least for the n semiconductor here assumed.

It is carried partly by electrons and partly by holes in this region of the current path. Introducing the mobility ratio

$$b = \frac{\mu_n}{\mu_p} \tag{V.2.21}$$

we find

$$e\mu_n n_0 Q(E_0 + E_1(x)) = I_c - e\mu_p[bn_1(x) + p_1(x)]QE_0 \tag{V.2.22}$$

The injected hole concentration p_1 decays with time according to the exponential law (V.2.15) by recombination. If we consider an arbitrarily chosen number of holes at a distance $L_c - x$ from the point I, the concentration $p(x)$ has decreased by a factor

$$e^{-t/\tau_p} = e^{-(L_c - x)/\tau_p v_{drift}}$$

compared with the original concentration $p(L_c)$, and a time

$$t = \frac{L_c - x}{v_{drift}}$$

has passed since the introduction of these holes into the n germanium. The same applies to the additional electron concentration $n_1(x)$ which has the same value as $p_1(x)$ for reasons of neutrality. We find therefore

$$n_1(x) = p_1(x) = p_1(L_c)\, e^{-\frac{L_c - x}{\tau_p v_{drift}}} \tag{V.2.23}$$

We substitute this in (V.2.22) and introduce the hole contribution of the collector current I_c at point I

$$e\mu_p p_1(L_c) E_0 \cdot Q = I_{c_p}(L_c) \tag{V.2.24}$$

and finally integrate (V.2.22) from $x = 0$ to $x = L_c$:

$$e\mu_n n_0 Q \int_{x=0}^{x=Lc} (E_0 + E_1(x))\, dx = I_c \cdot L_c - I_{c_p}(L_c) \cdot (b + 1) \int_{x=0}^{x=L_c} e^{-\frac{L_c - x}{\tau_p v_{drift}}}\, dx \tag{V.2.25}$$

Since $E_0 + E_1(x) = -\text{grad}\, V$, the field-strength integral on the left side leads to the value $-(U_I - U_c) = U_c - U_I$, i.e., the total voltage between the collector c and point I. Introducing the unmodulated "collector resistance,"

$$r_c = \frac{L_c}{e\mu_n n_0 Q} \tag{V.2.26}$$

dividing by L_c, and carrying out the integration on the right side of

(V.2.25) results in

$$\frac{1}{r_c}(U_c - U_I) = I_c - (1+b)I_{c_p}(L_c)\cdot\frac{\tau_p v_{drift}}{L_c}\left[1 - e^{-\frac{L_c}{\tau_p v_{drift}}}\right] \quad \text{(V.2.27)}$$

We introduce an abbreviation

$$\beta = \frac{\tau_p v_{drift}}{L_c}\left(1 - e^{-\frac{L_c}{\tau_p v_{drift}}}\right) \quad \text{(V.2.28)}$$

In terms of the transit time between point I and collector c

$$t_{tr} = \frac{L_c}{v_{drift}} \quad \text{(V.2.29)}$$

we obtain

$$\beta = \frac{\tau_p}{t_{tr}}\left(1 - e^{-\frac{t_{tr}}{\tau_p}}\right) \quad \text{(V.2.30)}$$

This "transport factor β" represents the fraction of the holes injected by the emitter at $x = L_c$ which actually arrives at the collector, i.e., which is not lost by recombination. From (V.2.27) we find then

$$\frac{1}{r_c}(U_c - U_I) = I_c - (1+b)\beta I_{c_p}(L_c) \quad \text{(V.2.31)}$$

Finally we can replace $I_{c_p}(L_c)$ by the negative hole contribution $-I_{e_p}$ of the emitter current or use the injection efficiency mentioned on page 120 so that

$$I_{c_p}(L_c) = -\gamma I_e \quad \text{(V.2.32)}$$

This equation is obtained by resolving Kirchhoff's law

$$I_b + I_e + I_c = 0 \quad \text{(V.2.33)}$$

into an equation for the electron currents

$$I_b + (1-\gamma)I_e + I_{c_n}(L_c) = 0 \quad \text{(V.2.34)}$$

and for the hole currents[1]

$$0 + \gamma I_e + I_{c_p}(L_c) = 0 \quad \text{(V.2.35)}$$

Equations (V.2.31) and (V.2.32) combine to yield

$$U_c - U_I = (1+b)\beta\gamma r_c I_e + r_c I_c \quad \text{(V.2.36)}$$

If we apply Ohm's law to the path between the base and point I and, further, use (V.2.33) after introducing the quantity

$$0 - U_I = +r_b I_b \quad \text{(V.2.37)}$$

[1] The hole contribution to the base current I_b is zero, because the field E_0 draws the injected holes away toward the collector on the right.

we obtain finally

$$\alpha_e = (1 + b)\beta\gamma \tag{V.2.38}$$

or

$$U_c - r_b(I_e + I_c) = \alpha_e r_c I_e + r_c I_c$$

$$U_c = (r_b + \alpha_e r_c) I_e + (r_b + r_c) I_c \tag{V.2.39}$$

This is the current-voltage relation in the *secondary* circuit of the filamentary transistor. The equation for the primary circuit can be obtained in a much simpler way.

Between the metal of the emitter point and the body of the germanium at point I, we find the voltage $U_e - U_I$ (see Fig. V.2.1). The relation between this voltage and the current I_e, i.e., the emitter characteristic, is not linear:

$$U_e - U_I = f(I_e) \tag{V.2.40}$$

Equations (V.2.37) and (V.2.33) yield for the current-voltage relation in the primary circuit of the filamentary transistor

$$U_e = r_b I_e + f(I_e) + r_b I_c \tag{V.2.41}$$

These equations are usually linearized by considering only small deviations u_e, u_c, i_e, i_c from the d-c operating point U_e, U_c, I_e, I_c. Introducing the differential emitter resistance

$$r_e = f'(I_e) \tag{V.2.42}$$

we find

$$u_e = (r_b + r_e) i_e + r_b i_c \tag{V.2.43}$$

$$u_c = (r_b + \alpha_e r_c) i_e + (r_b + r_c) i_c \tag{V.2.44}$$

e. The Amplification Properties of the Filamentary Transistor

Comparing (V.2.43), (V.2.44) with (V.2.18), (V.2.19), we obtain

$$r_{11} = r_b + r_e \qquad r_{12} = r_b$$

$$r_{21} = r_b + \alpha_e r_c \qquad r_{22} = r_b + r_c \tag{V.2.45}$$

With the help of (V.2.45) we can now evaluate the general amplification formulas (V.6.06), (V.6.07), and (V.6.08) for the special case of the filamentary transistor: Short-circuit current amplification ($R_L = 0$):

$$\left[\frac{i_c}{i_e}\right]_{\text{short circuit}} = -\frac{r_b + \alpha_e r_c}{r_b + r_c} \tag{V.2.46}$$

Open-circuit voltage amplification ($R_L = \infty$):

$$\left[\frac{u_L}{u_e}\right]_{\text{open circuit}} = -\frac{r_b + \alpha_e r_c}{r_b + r_e} \tag{V.2.47}$$

Power amplification under matched conditions ($R_L = r_{22} = r_b + r_c$):

$$\left[\frac{u_L i_c}{u_e i_e}\right]_{\text{matched}} = +\frac{1}{4}\frac{(r_b + \alpha_e r_c)^2}{(r_b + r_e)(r_b + r_c)}\frac{1}{1 - \frac{1}{2}\frac{(r_b + \alpha_e r_c) r_b}{(r_b + r_e)(r_b + r_c)}} \tag{V.2.48}$$

If we succeed in eliminating the base resistance r_b, which causes instabilities,[1] by proper structural design, i.e., by locating the emitter as far as possible to the left near the base, we obtain:

Short-circuit current amplification ($R_L = 0$):

$$\left[\frac{i_c}{i_e}\right]_{\text{short circuit}} = -\alpha_e \tag{V.2.49}$$

Open-circuit voltage amplification ($R_L = \infty$):

$$\left[\frac{u_L}{u_e}\right]_{\text{open circuit}} = -\alpha_e \frac{r_c}{r_e} \tag{V.2.50}$$

Power amplification under matched conditions ($R_L = r_{22} = r_c$):

$$\left[\frac{u_L i_c}{u_e i_e}\right]_{\text{matched}} = +\frac{1}{4}\alpha_e^2 \frac{r_c}{r_e} \tag{V.2.51}$$

If we finally add Eq. (V.2.38)

$$\alpha_e = (1 + b)\beta\gamma \tag{V.2.38}$$

with (V.2.21)

$$b = \frac{\mu_n}{\mu_p} \tag{V.2.21}$$

and with (V.2.30)

$$\beta = \frac{\tau_p}{t_{tr}}\left(1 - e^{-\frac{t_{tr}}{\tau_p}}\right) \tag{V.2.30}$$

we arrive at a very clear picture of the physical relations:

1. The largest possible injection efficiency γ, namely, unity, is advantageous for all three types of amplification. This is plausible since the emitter current i_e, which is controlled by the emitter voltage u_e, consists then entirely of modulating holes. On the other hand, all amplification properties are lost as $\gamma \to 0$. The conductance of the current path cannot be modulated by the injection of electrons into n germanium.

2. The largest possible transport factor β, namely, unity, is advantageous for all three types of amplification. This is plausible since all

[1] See in this connection p. 148. According to Eq. (V.2.43) we have $r_{12} = r_b$ for the filamentary transistor.

injected holes are then effective along the entire path from the emitter to the collector and are not lost by recombination before reaching the latter.

3. The largest possible mobility ratio b is advantageous for all three types of amplification. This is plausible since many electrons of high velocity, contributing correspondingly little to the space charge, are required to neutralize holes of low velocity, contributing greatly to the space charge. The current amplification which is attained with an ideal emitter ($\gamma = 1$) and ideal transport factor ($\beta = 1$) is often called the *true* current amplification α_i. It is in the case of the filamentary transistor

$$\alpha_i = 1 + b$$

and therefore depends here on the space-charge action of the holes.

In general, α_i gives the effectiveness of a hole arriving at the collector in modulating the collector current. α_i is, therefore, defined as the ratio of the *total* collector current variation i_c to the initiating variation i_{c_p} of the hole current arriving at the collector

$$\alpha_i = \left[\frac{i_c}{i_{c_p}}\right]_{u_c=0} \tag{V.2.52}$$

The secondary condition $u_c = 0$ assures that the collector current variation is caused only by i_{c_p}.

4. Let us consider the special case $\gamma = 1$, $\beta = 1$, $b = 0$. The emitter current thus consists entirely of modulating minority carriers which are not reduced by recombination but are all captured by the collector. They are assumed to be much faster than the majority carriers, so that they modulate the conductance of the path from emitter to collector only with their own conductivity and not with the conductivity of additional majority carriers needed for reestablishing neutrality.

In this case, the current amplification is unity, e.g., no current amplification takes place. However, voltage and power amplification are still obtained, if it is possible to make the collector resistance r_c large compared to the emitter resistance r_e—for instance, by means of a long and thin current path from emitter to collector.[1]

This special case represents, in a way, the pure and unadulterated transistor effect: Injected minority carriers modulate the conductance of the current path of a "local" battery merely by their presence.

[1] The difficulty in realizing such a condition lies in the fact that it is hard to keep the recombination small in spite of the length and the small cross section. A consequent reduction of β and hence α_e much below 1 would spoil everything.

This results in voltage and power amplification without current amplification. Current amplification is possible only if, in addition, the space-charge action of the minority carriers, requiring a concentration increase of the majority carriers, comes into play.

§3. The *n-p-n* Transistor

We have emphasized in the introduction, as a common principle of the mechanism of some transistor types, that the conductance of a current path containing a strong "local" current source is influenced by the small controlling power of a current coming from "far away." In the filamentary transistor the influenced current path is a piece of ohmic conductor, whereas in the *n-p-n* transistor[1] it is a reverse-biased *p-n* junction (see Fig. V.3.1). The cause of the blocking effect of such a *p-n* junction is the depletion of minority carriers in the diffusion tails, which ordinarily act as current sources.[2] Counteracting this carrier depletion by the injection of a varying number of minority carriers results in a control action on the reverse current of the junction. The injector or emitter in an *n-p-n* transistor is not a forward-biased point contact but another *p-n* junction which is biased in the forward direction, in contrast to the collector.[3] Thus we arrive at the *n-p-n* transistor shown in Fig. V.3.1 as the final stage. The operation of this transistor type is, therefore, based roughly on the following mechanisms: The left *n-p* junction, which is biased in the forward direction,

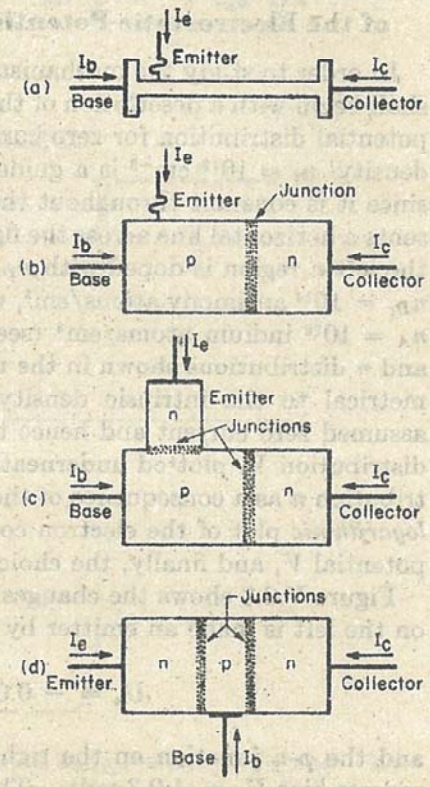

FIG. V.3.1. The transformation of the filamentary transistor into the *n-p-n* transistor.

[1] W. Shockley, *Bell System Tech. J.*, **28**: 435 (1949). W. Shockley, M. Sparks, and G. K. Teal, *Phys. Rev.*, **83**: 151 (1951).

[2] See pp. 102 to 103.

[3] See p. 106.

emits electrons into the central p layer; these electrons are collected by the p-n junction on the right which is biased in the reverse direction, and the reverse resistance of this p-n junction acting as collector is a sensitive function of the number of captured electrons. The *electron* currents flowing in the n-p-n transistor are therefore essential for its functioning. However, it will be necessary to consider also the hole components of the emitter and collector currents in the following quantitative treatment.

a. The Spatial Distribution of the Carriers and the Variation of the Electrostatic Potential

In order to study the mechanism of an n-p-n transistor in detail, we shall begin with a description of the concentration distribution and the potential distribution for zero current (see Fig. V.3.2). The intrinsic density[1] $n_i \approx 10^{13}$ cm^{-3} is a guide for the concentration distribution, since it is constant throughout the entire crystal and therefore represents a horizontal line across the figure. We assume that, for instance, the left n region is doped with $n_{Dl} = 10^{16}$ and the right n region with $n_{Dr} = 10^{14}$ antimony atoms/cm^3, whereas the central p region contains $n_A = 10^{15}$ indium atoms/cm^3 (see Fig. V.3.2).[2] This leads to the p and n distributions shown in the middle of Fig. V.3.2, which are symmetrical to the intrinsic density $n_i \approx 10^{13}$ cm^{-3} because we have assumed zero current and hence thermal equilibrium. The potential distribution V, plotted underneath, is identical with the electron distribution n as a consequence of the Boltzmann principle (IV.6.11), the *logarithmic* plot of the electron concentration n, the *linear* plot of the potential V, and finally, the choice of suitable scales.

Figure V.3.3 shows the changes which take place if the n-p junction on the left is made an emitter by applying forward bias

$$U_e = -0.078 \text{ volt} \approx -\frac{3kT}{e}$$

and the p-n junction on the right is made a collector by applying a reverse bias $U_c = +0.3$ volt. The concentration distributions within the transition regions of both p-n junctions maintain almost entirely the original Boltzmann character[3] in spite of the current flow.

The hole concentration p is therefore increased by the factor $e^{e|U_e|/kT} = e^{-eU_e/kT}$ at the point x_e and is decreased by the factor

[1] E. M. Conwell quotes for germanium at room temperature the more accurate value $n_i = 2.5 \cdot 10^{13}$ cm^{-3}. See *Proc. IRE*, 40: 1329 (1952), Table II.

[2] The subscripts l and r in n_{Dl} and n_{Dr} indicate left and right.

[3] See p. 100 and Fig. IV.8.1.

$e^{-e|U_c|/kT} = e^{-eU_c/kT}$ at the point x_c. The increase at the point x_c leads to a diffusion tail which decreases toward the left and in which p is reduced by the factor e for each diffusion length L_{pl}.[1] This diffusion tail determines the hole component of the emitter current. In the evaluation of the decrease at the point x_c, we must bear in mind the

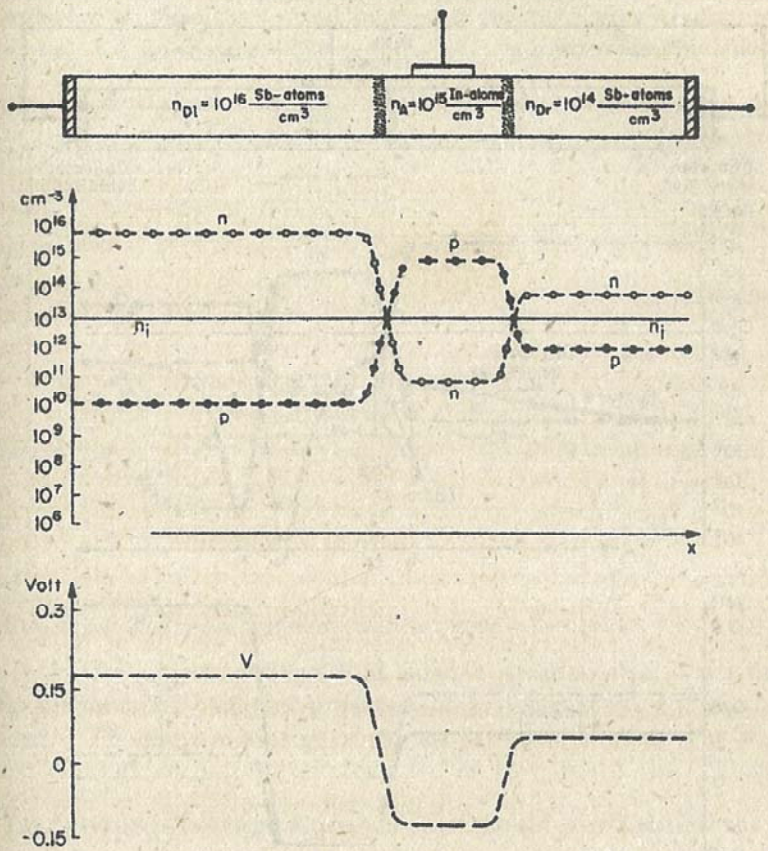

FIG. V.3.2. The *n-p-n* transistor. Concentration and potential distribution. The zero-current case.

logarithmic plot of p. The concentration decreases from the value p_{nr} to the value $p_{nr}\, e^{-eU_c/kT}$, which can here be regarded as practically equal to zero. This decrease from right to left takes place within a few diffusion lengths[2] L_{pr} and determines the hole component of the collector current.

[1] See p. 99. The subscripts l and r in L_{pl} and L_{pr} indicate again left and right.
[2] See Fig. IV.8.3, lower part.

The increase of the electron concentration n at the point x_{bl} leads to a diffusion tail which decays toward the right and extends in principle over several diffusion lengths L_n. The adjoining concentration decrease to the right, from the equilibrium density $n_p = n_i^2/n_A$ in the interior of the p region to practically zero at the point x_{br} extends, in

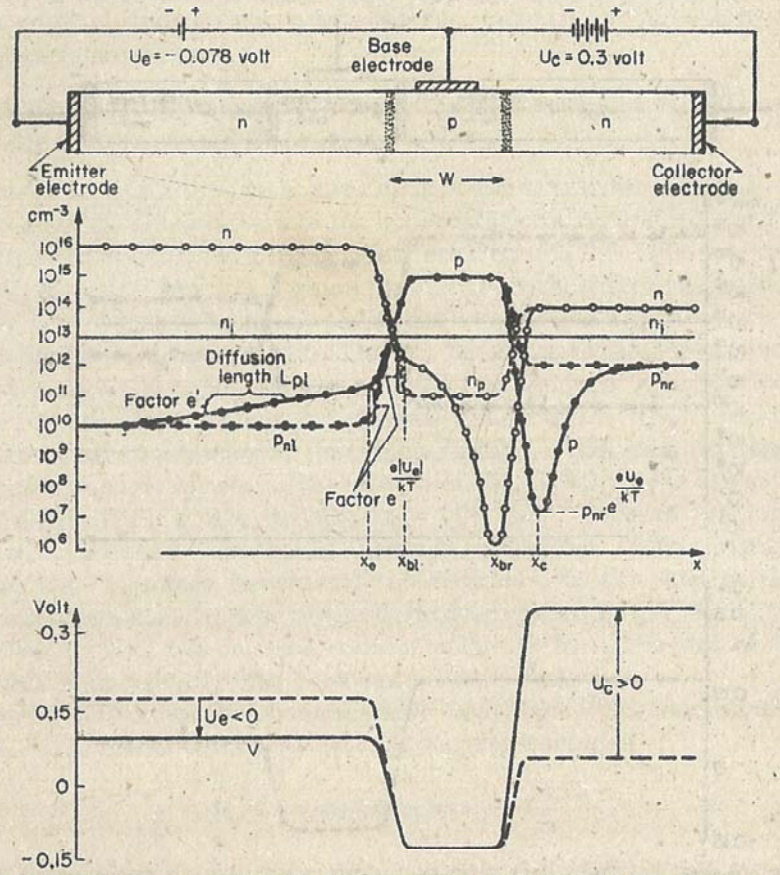

FIG. V.3.3. The *n-p-n* transistor. Concentration and potential distribution. The operating case.

principle, also over several diffusion lengths L_n. In Fig. V.3.3 we have made the assumption that the width W of the p layer is smaller than, or at most equal to, the diffusion length L_n, in contrast to Fig. V.3.4, so that the left and the right diffusion tails overlap and electrons are transported in a single diffusion process from the emitter at the left to the collector at the right. Thus the small electron supply from the p layer which is limited to the saturation-current value is enhanced

and modulated by the injection from the emitter. We see here already that this is possible only if the condition $W \leq L_n$ is fulfilled. If the p layer becomes too wide ($W \gg L_n$, Fig. V.3.4), the diffusion tail on the left drops to zero so that the electron concentration n remains horizontal for a certain distance at the thermal equilibrium value $n_p = n_i^2/n_A$-; the diffusion decay sets in beyond this point, in front of the collector junction. In the region of the p layer where n is now horizontal, we find essentially a pure hole current; for the electrons

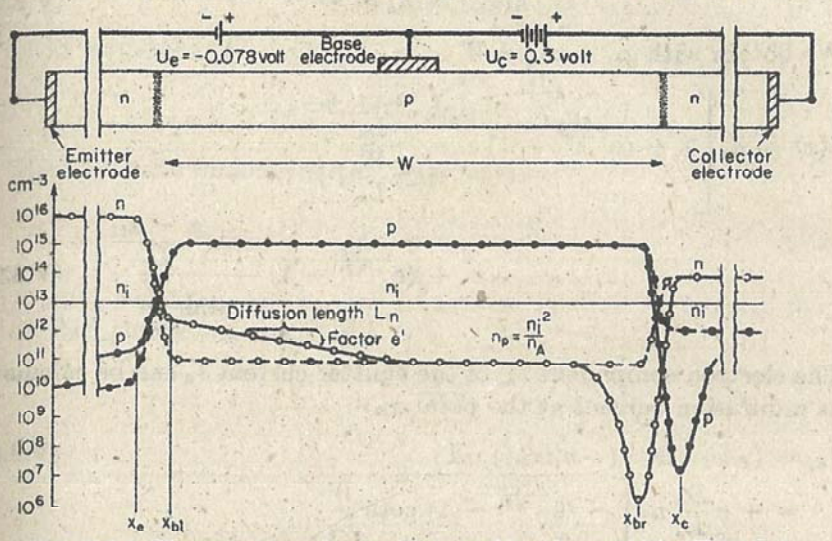

FIG. V.3.4. The *n-p-n* transistor with too wide a base width ($W \gg L_n$).

with their low concentration could furnish a substantial contribution to the current only in the presence of a nonexistent concentration gradient. The great majority of the electrons injected by the emitter are lost through recombination long before they reach the collector.

b. The Current-Voltage Equations of the *n-p-n* Transistor

The conditions in a diffusion tail were calculated in connection with the *p-n* rectifier. The fact that the divergence of the diffusion current is equal to the difference between recombination and pair formation leads to the differential equation

$$\frac{d^2}{dx^2}(n(x) - n_p) - \frac{1}{L_n^2}(n(x) - n_p) = 0 \qquad \text{(IV.8.03)}$$

We were then satisfied with the particular solution $A\,e^{+x/L_n}$ which corresponds to an electron current vanishing at minus infinity. Now we require for the description of the electron concentration in the

p layer of our n-p-n transistor the complete solution

$$n(x) - n_p = A\,e^{+\frac{x}{L_n}} + B\,e^{-\frac{x}{L_n}} \tag{V.3.01}$$

The integration constants A and B are determined by the boundary conditions

$$n(x_{bl}) = n_p\,e^{-\frac{eU_e}{kT}} \tag{V.3.02}$$

$$n(x_{br}) = n_p\,e^{-\frac{eU_c}{kT}} \tag{V.3.03}$$

We obtain with $x_{br} - x_{bl} \approx W$

$$n(x) = n_p\left[1 + (e^{-\frac{eU_e}{kT}} - 1)\frac{\sinh\frac{x_{br} - x}{L_n}}{\sinh\frac{W}{L_n}} + (e^{-\frac{eU_c}{kT}} - 1)\frac{\sinh\frac{x - x_{bl}}{L_n}}{\sinh\frac{W}{L_n}}\right] \tag{V.3.04}$$

The electron component I_{e_n} of the emitter current I_e can be calculated as a diffusion current at the point x_{bl}:

$$I_{e_n} = (-e)\cdot D_n\cdot(-n'(x_{bl}))\cdot A \tag{V.3.05}$$

$$= +e\frac{D_n}{L_n}n_p\left[-(e^{-\frac{eU_e}{kT}} - 1)\coth\frac{W}{L_n} + (e^{-\frac{eU_c}{kT}} - 1)\frac{1}{\sinh\frac{W}{L_n}}\right]\cdot A \tag{V.3.06}$$

A is here the cross-sectional "area" of the n-p-n transistor.

If we use the Nernst-Townsend-Einstein relation $D_n = \mu_n kT/e$ in (V.3.06) and define the following conductances

$$e\mu_n n_p\frac{A}{L_n}\coth\frac{W}{L_n} = G_{ll_n} \tag{V.3.07}$$

$$-e\mu_n n_p\frac{A}{L_n}\frac{1}{\sinh\frac{W}{L_n}} = G_{lr_n} \tag{V.3.08}$$

we find for the electron component I_{e_n} of the emitter current I_e

$$I_{e_n} = -G_{ll_n}\frac{kT}{e}(e^{-\frac{eU_e}{kT}} - 1) - G_{lr_n}\frac{kT}{e}(e^{-\frac{eU_c}{kT}} - 1) \tag{V.3.09}$$

A hole component must be added; for a hole diffusion tail extends into the left n region

$$p(x) = p_{nl} + p_{nl}\,(e^{-\frac{eU_e}{kT}} - 1)\,e^{\frac{x-x_e}{L_{pl}}} \qquad \text{(V.3.10)}$$

At the beginning $x = x_e$ of the diffusion tail there is a diffusion current

$$I_{e_p} = (+e)\cdot D_p\cdot(-p'(x_e))\cdot A$$

$$= -e\mu_p p_{nl}\cdot\frac{A}{L_{pl}}\cdot\frac{kT}{e}\cdot(e^{-\frac{eU_e}{kT}} - 1) \qquad \text{(V.3.11)}$$

or in simplified form

$$I_{e_p} = -G_{ll_p}\cdot\frac{kT}{e}\,(e^{-\frac{eU_e}{kT}} - 1) \qquad \text{(V.3.12)}$$

Here we have introduced the "conductance"

$$G_{ll_p} = e\mu_p p_{nl}\frac{A}{L_{pl}} \qquad \text{(V.3.13)}$$

Finally we obtain for the emitter current $I_e = I_{e_n} + I_{e_p}$ with the help of Eqs. (V.3.09) and (V.3.12)

$$I_e = +G_{ll}\frac{kT}{e}\,(1 - e^{-\frac{eU_e}{kT}}) + G_{lr}\frac{kT}{e}\,(1 - e^{-\frac{eU_c}{kT}}) \qquad \text{(V.3.14)}$$

and correspondingly for the collector current

$$I_c = +G_{rl}\frac{kT}{e}\,(1 - e^{-\frac{eU_e}{kT}}) + G_{rr}\frac{kT}{e}\,(1 - e^{-\frac{eU_c}{kT}}) \qquad \text{(V.3.15)}$$

with the conductances

$$G_{ll} = G_{ll_n} + G_{ll_p} = e\mu_n n_p\frac{A}{L_n}\coth\frac{W}{L_n} + e\mu_p p_{nl}\frac{A}{L_{pl}} \qquad \text{(V.3.16)}$$

$$G_{lr} = G_{lr_n} = -e\mu_n n_p\frac{A}{L_n}\frac{1}{\sinh\frac{W}{L_n}} \qquad \text{(V.3.17)}$$

$$G_{rl} = G_{rl_n} = -e\mu_n n_p\frac{A}{L_n}\frac{1}{\sinh\frac{W}{L_n}} \qquad \text{(V.3.18)}$$

$$G_{rr} = G_{rr_n} + G_{rr_p} = e\mu_n n_p\frac{A}{L_n}\coth\frac{W}{L_n} + e\mu_p p_{nr}\frac{A}{L_{pr}} \qquad \text{(V.3.19)}$$

If we replace I_e in (V.3.14) and (V.3.15) by $I_e + i_e$, I_c by $I_c + i_c$, U_e by $U_e + u_e$, and U_c by $U_c + u_c$, and if we expand on the right side the small quantities u_e and u_c and make use of (V.3.14) and (V.3.15), we find for the small variations i_e, u_e, . . . about the operating point

I_e, U_e, . . . the equations

$$i_e = G_{ll}\, e^{-\frac{eU_e}{kT}} u_e + G_{lr}\, e^{-\frac{eU_c}{kT}} u_c = g_{11} u_e + g_{12} u_c \qquad \text{(V.3.20)}$$

$$i_c = G_{rl}\, e^{-\frac{eU_e}{kT}} u_e + G_{rr}\, e^{-\frac{eU_c}{kT}} u_c = g_{21} u_e + g_{22} u_c \qquad \text{(V.3.21)}$$

Thus we have found the current-voltage characteristics of the *n-p-n* transistor in the conductance form, whereas for the filamentary transistor we arrived at the resistance form [(V.2.43) and (V.2.44)].

c. The Transport Factor β and the Efficiencies γ_e and γ_c for the *n-p-n* Transistor

The foregoing equations allow us to obtain a more accurate quantitative picture of our previous qualitative concepts of the mechanism of the *n-p-n* transistor. The emitter has the task of emitting toward the collector a number of electrons proportional to the emitter-voltage variation u_e. This task is better fulfilled, the more *electrons* are contained in the emitter current resulting from u_e alone

$$[i_e]_{u_c=0} = g_{11} u_e = (G_{ll_n} + G_{ll_p})\, e^{-\frac{eU_e}{kT}} u_e$$

A measure of merit of the emitter is the injection efficiency

$$\gamma_e = \frac{[i_{e_n}]_{u_c=0}}{[i_e]_{u_c=0}} = \frac{G_{ll_n}}{G_{ll_n} + G_{ll_p}} = \frac{G_{ll_n}}{G_{ll}} \qquad \text{(V.3.22)}$$

We shall later find it useful to introduce a corresponding collection efficiency for the collector

$$\gamma_c = \frac{G_{rr_n}}{G_{rr}} \qquad \text{(V.3.23)}$$

The effectiveness of the emitter is not assured solely by an adequate content of electrons. The emitted electrons must also be collected by the collector and must not be lost by recombination. A measure of the proper functioning of the collector is, therefore, the ratio of the electron current[1] $[-i_{c_n}]_{u_c=0}$ arriving at the collector and the electron current $[i_{e_n}]_{u_c=0}$ emitted by the emitter. The transport factor β is thus given by

$$\beta = \frac{[-i_{c_n}]_{u_c=0}}{[i_{e_n}]_{u_c=0}} = \frac{-G_{rl_n}}{G_{ll_n}} = \frac{1}{\cosh \frac{W}{L_n}} \qquad \text{(V.3.24)}$$

[1] The minus sign is required because for i_c the positive direction is defined oppositely to that for i_e.

We see now that

$$\beta = \frac{1}{\cosh \frac{W}{L_n}} \leq 1 \tag{V.3.25}$$

and that the optimum value 1 is dependent on $W \ll L_n$ as we found on page 131 in a qualitative manner. The differential conductances $g_{11} \cdots g_{22}$ can be described with the help of γ_e, γ_c, and β, as we can see from Eqs. (V.3.16) to (V.3.25):

$$\left.\begin{aligned} g_{11} &= +\frac{1}{\beta\gamma_e} G \cdot e^{-\frac{e}{kT}U_e} & g_{12} &= - G \cdot e^{-\frac{e}{kT}U_c} \\ g_{21} &= - G \cdot e^{-\frac{e}{kT}U_e} & g_{22} &= +\frac{1}{\beta\gamma_c} G \cdot e^{-\frac{e}{kT}U_c} \end{aligned}\right\} \tag{V.3.26}$$

For brevity we use the substitution

$$G = e\mu_n n_p \frac{A}{L_n} \frac{1}{\sinh \frac{W}{L_n}} \tag{V.3.27}$$

d. Current and Voltage Amplification of the *n-p-n* Transistor

We gain further insight into the mechanism of the *n-p-n* transistor if we derive the current amplification for the short-circuited collector ($u_c = 0$) from the current-voltage equations (V.3.20) to (V.3.21) and the conductances (V.3.26):

$$\left[\frac{i_c}{i_e}\right]_{u_c=0} = \frac{g_{21}}{g_{11}} = -\beta \cdot \gamma_e \tag{V.3.28}$$

The current amplification can at best reach the value 1 when the transport factor β and the injection efficiency γ_e of the emitter have their optimum values 1. This requirement for maximum current amplification is thus the same as for the filamentary transistor. However, in contrast to Eqs. (V.2.49) and (V.2.38) we do not have the favorable factor $1 + b = +\mu_n/\mu_p$ in addition to $\beta\gamma_e$. The *true* current amplification α_i is, therefore, equal to 1, as can also be seen from a suitable modification of the defining equation (V.2.52)

$$\alpha_i = \left[\frac{i_c}{i_{c_n}}\right]_{u_c=0} \tag{V.3.281}$$

provided that the secondary current-voltage equation (V.3.21) and the fact that $G_{rl} = G_{rl_n}$ (V.3.18) are utilized in the evaluation of (V.3.281).

$\alpha_i = 1 + b$ for the filamentary transistor because of the space-charge compensation of the injected minority carriers by additional majority

carriers. Such an increase of the majority carrier concentration exists also in the *n-p-n* transistor. However, it does not affect the current as long as the latter is determined by the yield of the diffusion tails. Only when the ohmic resistance of the path becomes important, namely, at high currents, may a similar effect be expected in the *n-p-n* transistor.[1]

The voltage amplification with open-circuited collector ($i_c = 0$) is, according to (V.3.21) and (V.3.26),

$$\left[\frac{u_c}{u_e}\right]_{i_c=0} = -\frac{g_{21}}{g_{22}} = +\beta\gamma_c\, e^{+\frac{e}{kT}(U_c - U_e)} \qquad \text{(V.3.29)}$$

We have now $U_c > 0$ and $U_e < 0$, and the difference is

$$U_c - U_e \gg \frac{kT}{e} = 26 \text{ mv} > 0$$

so that the exponential factor is large compared with 1. Even if β and γ_c do not have their optimum value, large voltage amplifications are obtained. In the *n-p-n* transistors described by Shockley, Sparks, and Teal[2] the conductivity of the collector *n* layer is, incidentally, about an order of magnitude smaller than that of the *p* layer with the base electrode. Accordingly, the collector current consists predominantly of holes. Despite a γ_c appreciably smaller than 1, these transistors exhibit high voltage amplification in accordance with (V.3.29). However, apart from the practical importance of the case $\gamma \ll 1$, the discussion of this condition is informative.

e. The Special Case[3] $\gamma_c \ll 1$, $\gamma_e = 1$, $\beta = 1$

With $\gamma_e = 1$ and $\beta = 1$, the two current-voltage equations (V.3.20) and (V.3.21), utilizing (V.3.26), can be written in the form

$$i_e = G \cdot e^{-\frac{e}{kT}U_e} u_e - G \cdot e^{-\frac{e}{kT}U_c} u_c \qquad \text{(V.3.30)}$$

$$i_c = -G \cdot e^{-\frac{e}{kT}U_e} u_e + \frac{1}{\gamma_c} G \cdot e^{-\frac{e}{kT}U_c} u_c \qquad \text{(V.3.31)}$$

The feedback term $G \cdot e^{-\frac{e}{kT}U_c} u_c$ can now be omitted in (V.3.30).

[1] Early shows that, in weakly doped collectors, the minority carrier current is a field current and not a diffusion current. Thus *inherent* current amplification becomes possible. See J. M. Early, *Bell System Tech. J.*, **32**: 1271 (1953), particularly pp. 1306ff.

[2] W. Shockley, M. Sparks, and G. K. Teal, *Phys. Rev.*, **83**: 151 (1951).

[3] To bring out the principles involved, we assume that $\gamma_e = 1$ and $\beta = 1$ in addition to $\gamma_c \ll 1$.

Unlike the term below it, it contains the very small factor $e^{-\frac{e}{kT}U_c}$ uncompensated by $1/\gamma_c$, which, according to our assumptions, is large. The current-voltage equations then assume the form

$$i_e = G \cdot e^{-\frac{e}{kT}U_e} u_e$$
$$i_c = -G \cdot e^{-\frac{e}{kT}U_e} u_e + \frac{1}{\gamma_c} G \cdot e^{-\frac{e}{kT}U_c} u_c = -i_e + \frac{1}{\gamma_c} G \cdot e^{-\frac{e}{kT}U_c} u_c$$

By solving for the voltages we obtain the resistance form:

$$u_e = \frac{1}{G} e^{+\frac{e}{kT}U_e} i_e \qquad \text{(V.3.32)}$$

$$u_c = \gamma_c \frac{1}{G} e^{+\frac{e}{kT}U_c} i_e + \gamma_c \frac{1}{G} e^{+\frac{e}{kT}U_c} i_c \qquad \text{(V.3.33)}$$

These equations have now become identical with the current-voltage equations (V.2.43) and (V.2.44) of the filamentary transistor if they are written for the special case $r_b = 0$, $\beta = 1$, $\gamma = 1$, $b = 0$ discussed on page 124

$$u_e = r_e i_e \qquad \text{(V.3.34)}$$
$$u_c = r_c i_e + r_c i_c \qquad \text{(V.3.35)}$$

In fact, we can now see an almost complete analogy between the filamentary transistor and the *n-p-n* transistor:

1. In both cases an ideal emitter ($\gamma = 1$ or $\gamma_e = 1$) injects minority carriers which are collected without loss by the collector ($\beta = 1$).

2. A space-charge action and with it a *true* current amplification α_i is absent in both cases, in the filamentary transistor because of the somewhat artificial assumption $b = \mu_n/\mu_p \approx 0$ and in the *n-p-n* transistor because of the fundamental reasons discussed on page 135.

3. The injected minority carriers influence the conductance of a current path whose conductivity, without injection, is predominantly determined by carriers of the opposite polarity. This is, in contrast to point 2, a rather natural assumption for the filamentary transistor, whereas it seems somewhat artificial for the *n-p-n* transistor.[1]

This makes a detailed discussion of the opposite limiting case $\gamma_c = 1$ desirable, because characteristic features in the mechanism of

[1] Inasmuch as (V.3.29) shows that for high amplification γ_c ought to be as large as possible, namely, $\gamma_c \to 1$. Then the entire collector current consists of electrons which, in contrast to the holes, are subject to modulation. The fact that real transistors can, nevertheless, have $\gamma_c \ll 1$ (see pp. 135 to 136) is probably a consequence of the particular method of preparation.

the *n-p-n* transistor will appear which have no analogies in the filamentary transistor.

f. The Special Case $\gamma_c = 1$, $\gamma_e = 1$, $\beta = 1$

The current-voltage equations (V.3.30) and (V.3.31) assume in this case the form

$$i_e = G \cdot e^{-\frac{e}{kT}U_e} u_e - G \cdot e^{-\frac{e}{kT}U_c} u_c \qquad \text{(V.3.36)}$$

$$i_c = -G \cdot e^{-\frac{e}{kT}U_e} u_e + G \cdot e^{-\frac{e}{kT}U_c} u_c \qquad \text{(V.3.37)}$$

and we see that the base current is zero, independently of u_e and u_c:

$$i_b = -(i_e + i_c) \equiv 0 \qquad \text{(V.3.38)}$$

This particular case exhibits a certain analogy between the *n-p-n* transistor and the vacuum tube. Thus the base corresponds to the control grid by way of the absence of current, the emitter corresponds to

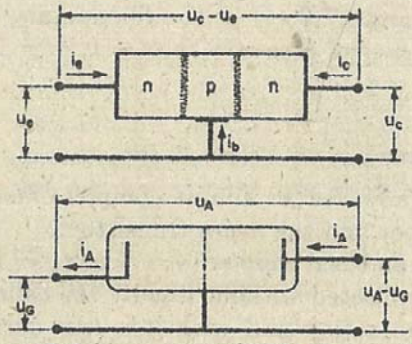

FIG. V.3.5. Comparison of the *n-p-n* transistor in the special case $\alpha_c = 1$, $\alpha_e = 1$, $\beta = 1$ with the vacuum tube.

the cathode by way of emission, and the collector corresponds to the anode by way of collection (see Fig. V.3.5).

If we write (V.3.37) in the form

$$i_c = G[e^{-\frac{e}{kT}U_e} - e^{-\frac{e}{kT}U_c}] \left\{(-u_e) + \frac{e^{-\frac{e}{kT}U_c}}{e^{-\frac{e}{kT}U_e} - e^{-\frac{e}{kT}U_c}} (u_c - u_e)\right\}$$

we see with the aid of Fig. V.3.5 and by comparison with the vacuum-tube equation[1]

$$i_A = S\{u_G + Du_A\}$$

[1] i_A = plate current, u_A = plate voltage, u_G = grid voltage, S = transconductance, D = reciprocal voltage amplification.