THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXXIII JULY 1954 NUMBER 4

Negative Resistance Arising from Transit Time in Semiconductor Diodes

By W. SHOCKLEY

(Manuscript received January 22, 1954)

The structural simplicity of two-terminal compared to three-terminal devices indicates the potential importance of two terminal devices employing semiconductors and having negative resistance at frequencies properly related to the transit time of carriers through them. Such negative resistances may be combined with unsymmetrically transmitting components, such as gyrators or Hall effect plates, to form dissected amplifiers *that may be made to simulate conventional three-terminal amplifiers and operate at high frequencies. The characteristics of several structures are analyzed on the basis of theory and it is found that negative resistances are possible for properly designed structures.*

1. NEGATIVE RESISTANCE AND DISSECTED AMPLIFIERS

Because the drift velocities of current carriers in semiconductors are smaller than the velocities attainable in vacuum tubes, transistor structures must be smaller to achieve comparable frequencies. In principle it is possible, of course, to make compositional structures (i.e., distributions of donors and acceptors) in semiconductor crystals on a scale much smaller than is possible for vacuum tubes. At present, however, the available techniques are limited and it may require many years before the ultimate potentialities are approached.

It is instructive, however, to speculate on some of these ultimate potentialities. For example a grain boundary formed of edge type dislocations is in a sense an analogue of a grid. Possibly it can be made into a grid by acting as a locus for an atmosphere of donors or acceptors. Evidently such a grid will approach the smallest spacing that can be achieved with any known form of matter. If the spacings perpendicular to the grid are made comparable to a mean free path of the carriers used, the device will operate like a vacuum tube with carrier velocities controlled by inertia rather than by mobility. It is not easy to conceive of a structure having the potentiality of operating at higher frequencies.

It is evident that the difficulty of making small structures increases with the number of electrodes. For example, it is now possible to make diodes which give usable rectification at frequencies above 10^{10} cps. In these the "working volume" is a very thin layer under the metal point. The thickness of this layer is controlled by surface treatments and the applied voltages. The diameter of the point, which is the minimum dimension mechanically controlled, is much larger than this thickness of the layer. In order to make a transistor of comparable frequency, it would be necessary to make structural elements having dimensions comparable to the thickness of the layer and this would be a much more exacting task than making the diode.

These considerations point out the importance of giving serious consideration to two-terminal structures as amplifying elements. It is possible, in principle at least, to have structures which are much smaller in one dimension than the other two and which exhibit negative resistance and thus give ac power at frequencies comparable to the reciprocal of the transit time across the small dimension.

The attractiveness of such negative resistance diodes for amplification is enhanced by the possibility of using them in *dissected amplifiers*[1.1] in combination with nonreciprocal elements such as gyrators or Hall effect plates. Combinations of negative resistance elements and nonreciprocal elements can lead to structures having gain and unsymmetrical transmission that simulate conventional amplifiers. The adjective *dissected* has been suggested for them since elements giving power gain are physically separated from those giving one-way transmission.

In this article we shall not consider the possible forms of dissected amplifiers, of which there are a wide variety. Instead we shall give an introductory treatment of some forms of negative resistance that may arise from transit time effects. In some cases the most instructive way of treating the structure is by way of the "impulsive impedance" and we devote most of the next section to considering this method.

2. THE IMPULSIVE IMPEDANCE AND NEGATIVE RESISTANCE

The impulsive impedance $D(t)$ for a two terminal device is defined in terms of its transient response to an impulse of current. Thus if the current through the device is

$$J(t) = J + j(t),$$

where J is the dc current and

$$j(t) = 0$$

except very near $t = 0$ and

$$\int j(t)\, dt = \delta Q,$$

then the voltage is

$$V(t) = V + v(t),$$

where V is the dc voltage and

$$v(t) = \delta Q\ D(t).$$

In other words, if in addition to the dc biasing current, a charge δQ is instantaneously forced through the circuit at time $t = 0$, the added voltage is $D(t)$ per unit charge. These equations also serve to introduce the notation used in this article:

Notation. In general, quantities that are functions of time or position will have the functional dependence explicitly indicated. In Sections 4 and 5, however, the symbol δ will be used to distinguish the transient parts δE and $\delta\rho$ from the dc parts of the electric field and charge density.

Capital $V(t)$ and $J(t)$ stand for total voltages and current. Without functional dependence upon (t) they are the dc parts. Similarly $v(t)$ and $j(t)$ are the ac or transient parts. A sinusoidal disturbance is represented by

$$v(t) = v \exp i\omega t,$$

$$j(t) = j \exp i\omega t.$$

where v and j are not functions of time. Where it is necessary to distinguish the displacement current at a particular location from the conduction current, as in the next section, we shall write

$$j(D,\ S_2,\ t),$$

meaning the displacement current across space charge region S_2 as a function of time.

In this section we shall treat J and j as circuit currents. In subsequent sections, we shall be concerned with current densities and shall use the same symbols.

The complex impedance of the device is evidently

$$Z(\omega) = v/j,$$

where v and j are the coefficients in the sinusoidal case.

In terms of the system of notation introduced above, $Z(\omega)$ may also be expressed in terms of $D(t)$ by expressing $j \exp i\omega t$ in terms of increments of charge

$$dQ = je^{i\omega t}\, dt,$$

and summing over all increments up to time t. This leads to

$$Z(\omega) = \int_0^\infty D(t) \exp(-i\omega t)\, dt.$$

A negative resistance will occur if

$$0 > \int_0^\infty D(t) \cos \omega t\, dt = (-1/\omega) \int_0^\infty D'(t) \sin \omega t\, dt,$$

the latter form coming from integration by parts for the case of $D(\infty) = 0$, the only situation treated in this article.

The use of $D(t)$ in analysing the potential merits of diode structures from the point of view of negative resistance is illustrated in Fig. 2.1. Here three cases of $D(t)$ together with certain cosine waves are shown. It is seen for case (a) that a negative real part of Z will be obtained. For case (b), the real part of Z is zero for the frequency shown; this represents a limit; for other frequencies, a positive real part will be ob-

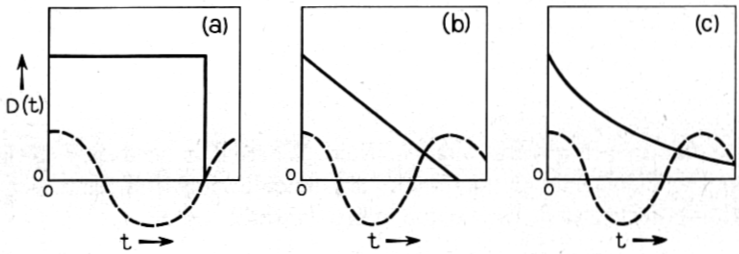

Fig. 2.1 — Some hypothetical $D(t)$ characteristics.

tained. Case (c) represents an exponential fall such as might occur for a capacitor and resistor in parallel. We shall discuss this example below.

The conclusions regarding (a) and (b) may be somewhat more easily seen from the corresponding $-D'(t)$ plots shown in Fig. 2.2. From part (a) it can be seen that the negative maximum in the sine wave at the end of the rectangular $D(t)$ plot is particularly favorable. From part (b) it is seen that no choice of ω will result in more negative area of sine wave than positive. For (c) it is evident that each positive half cycle of the sine wave gives a larger contribution than the following negative half cycle and hence that a positive resistance will be obtained.

For case (c), it is instructive to obtain the value of Z analytically by using

$$D(t) = C^{-1} \exp (-t/RC).$$

This leads correctly to

$$Z(\omega) = (R^{-1} + i\omega C)^{-1}.$$

For small values of ωRC, Z reduces to R; furthermore, for this case, the decay of $D(t)$ occurs while $\cos \omega t = 1$. Under these conditions

$$Z(\omega) = \int_0^\infty D(t)\, dt.$$

This result is useful for estimating the effect of quickly decaying contributions to $D(t)$. These evidently contribute a positive resistance to Z equal to the area under the $D(t)$ curve.

From these considerations it follows that an upward deviation from the linear fall in Fig. 2.1(b) towards Fig. 2.1(a) will result in negative resistance. In Sections 4 and 5 we shall see how particular structures may lead to such favorable, convex-upwards characteristics for $D(t)$.

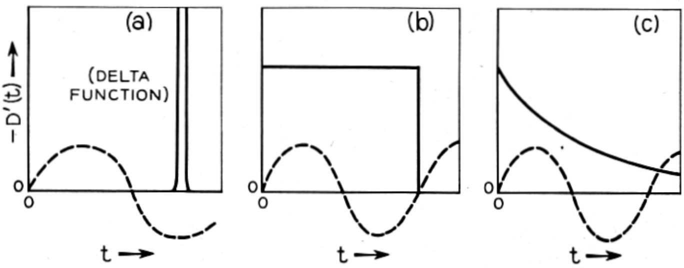

Fig. 2.2 — The $-D'(t)$ characteristic corresponding to fig. 2.1.

3. MINORITY CARRIER DELAY DIODE

As a first example we shall consider the behavior of the device shown in Fig. 3.1. We have chosen a p-n-p structure rather than an n-p-n so as to deal with positively charged carriers and thus avoid numerous minus signs in the equations. In this figure we have used capital letters P and N to designate specific regions, reserving the small letters to indicate carrier densities and conductivity types.

Several features that simplify the theoretical treatment should be noted:

(a) The P_1N junction is 100-fold more heavily doped on the P_1-side.

(b) The doping in the layer N varies exponentially with distance by a factor of 10 across the layer.

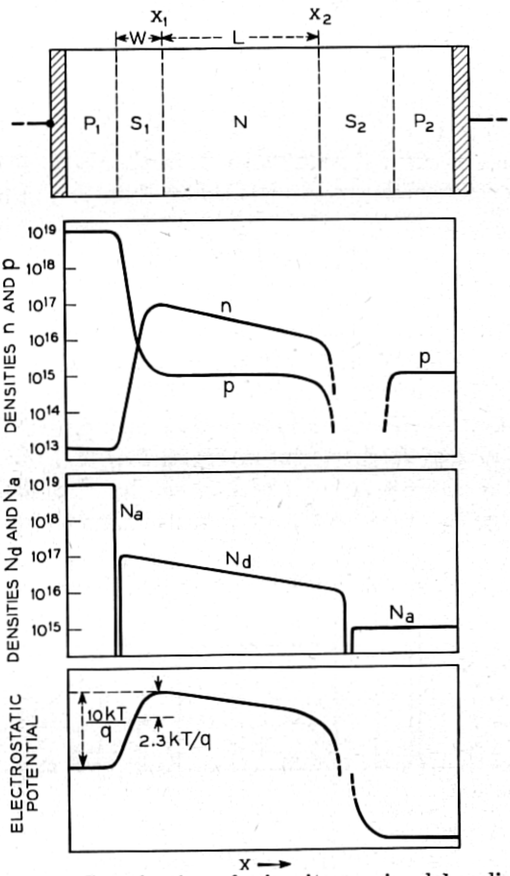

Fig. 3.1 — Constitution of minority carrier delay diode.

(c) Throughout N the concentration of holes is less by a factor of 10 than the electron concentration.

(d) The thickness of N is large compared to the depth of space charge penetration into it.

(e) The voltage drop across the space charge region S_2 is large compared to the other voltage drops.

The conditions lead at the operating frequency to the following consequences:

(1) The current across the first junction is carried preponderantly by holes.

(2) The hole drift in N is substantially unaffected by the ac field and thus represents the delayed diffusing and drifting current injected across the first junction.

(3) The ac voltage drop occurs chiefly across S_2.

We shall show below (Section 3.2) how (a) to (e) lead to (1) to (3), but we shall first bring out the importance of (1) to (3) by using them to give a simple treatment of the impedance of the diode.

3.1. *Calculation of Impedance.*

If the total current is

$$J(t) = J + je^{i\omega t}, \tag{3.1}$$

then the ac hole current across S_1 is also in the notation discussed in Section 2 with the addition of the symbol p to indicate holes

$$j(p, S_1, t) = je^{i\omega t}. \tag{3.2}$$

This current flows through the n-layer unaffected by the ac field and arrives at S_2 delayed and attenuated by a complex factor

$$\beta = |\beta| \exp(-i\theta). \tag{3.3}$$

Because of the high field in S_2, the transit time there is negligible so that the hole current arriving at P_2 is

$$j(p, S_2)e^{i\omega t} = \beta je^{i\omega t}. \tag{3.4}$$

In addition to this current, there is a dielectric displacement current in S_2 which is converted to hole conduction current in P_2. If the voltage drop across S_2 is

$$V(S_2, t) = V(S_2) + v(S_2)e^{i\omega t}, \tag{3.5}$$

then the ac displacement current is

$$j(D, S_2)e^{i\omega t} = i\omega C_2 v(S_2)e^{i\omega t}. \tag{3.6}$$

Now the total current is constant through the device, hence

$$j = j(p, S_2) + j(D, S_2) \tag{3.7}$$

which leads to

$$j = i\omega C_2 v(S_2)/(1 - \beta). \tag{3.8}$$

If $v(S_2)$ is substantially equal to the ac voltage across the unit, then the impedance is

$$\begin{aligned} Z &= v(S_2)/j = (1/i\omega C_2) + i\beta/\omega C_2 \\ &= (1/i\omega C_2) + (1/\omega C_2) \mid \beta \mid \exp i[(\pi/2) - \theta]. \end{aligned} \tag{3.9}$$

Evidently if $\theta > \pi$ and $\theta < 2\pi$, the second term will have a negative real part so that the diode will act as a power source.

If we neglect the ac electric field in N then β may be calculated in terms of the thickness $L = x_2 - x_1$ of the layer and the potential drop across the layer. This latter arises from the concentration ratio N_{d1}/N_{d2} between the two sides of N. Since the donor charge density is neutralized substantially entirely by electrons, and since almost no electron current flows, the electron concentration difference must result from a Boltzman factor (at $10^{17}/\text{cm}^3$ Fermi-Dirac statistics are not needed) and this leads to

$$\Delta V_n = (kT/q)ln(N_{d1}/N_{d2}) \tag{3.10}$$

for the potential drop across N. In N the electric field is thus

$$E = \Delta V_n/L. \tag{3.11}$$

The differential equation for hole concentration for a disturbance of frequency ω is

$$\dot{p} = i\omega p = -\mu p E \frac{\partial p}{\partial x} + D_p \frac{\partial^2 p}{\partial x^2}. \tag{3.12}$$

This linear differential equation has two linearly independent solutions. These must satisfy at x_2, the left edge of the space charge layer S_2, the boundary condition that the hole density is practically zero.[3.1] The appropriate solution is

$$p = e^{i\omega t} [e^{k_1(x-x_2)} - e^{k_2(x-x_2)}], \tag{3.13}$$

where

$$Lk_1 \equiv (x_2 - x_1)k_1 = \alpha[1 + (1 + i\gamma)^{1/2}], \tag{3.14}$$

$$Lk_2 \equiv (x_2 - x_1)k_2 = \alpha[1 - (1 + i\gamma)^{1/2}] \tag{3.15}$$

where

$$\alpha = q\Delta V/2kT, \tag{3.16}$$

$$\gamma = 4\omega D_p/u^2, \tag{3.17}$$

$$u = \mu_p E = \mu_p \Delta V/L. \tag{3.18}$$

The current is

$$j(p, x, t) = q(up - D_p \partial p/\partial x) \tag{3.19}$$

and the ratio of currents at x_1 and x_2, which is β by definition, is

$$\beta = j(p, x_2, t)/(j(p, x_1, t),$$

$$= \frac{Lk_1 - Lk_2}{Kk_1 \exp(-Lk_2) - Kk_2 \exp(-Lk_1)} \tag{3.20}$$

$$= \frac{2(1 + i\gamma)^{1/2} e^{\alpha}}{[1 + (1 + i\gamma)^{1/2}] \exp \alpha(1 + i\gamma)^{1/2} - [1 - (1 + i\gamma)^{1/2}] \exp - \alpha(1 + i\gamma)^{1/2}}.$$

The phase lag in β must exceed 180° or π in order to give negative resistance. It can be seen that this phase factor must result from the first exponential in the denominator by the line of reasoning suggested below: The real part of the exponent is larger than the imaginary part. Hence the absolute ratio of the two exponentials is at least 2π. For this condition the second term in the denominator is negligible compared to the first. Hence the phase of (3.20) is determined largely by the first exponential. As a helpful approximation we may neglect the second term and write

$$\beta \doteq \frac{2(1 + i\gamma)^{1/2} \exp[\alpha - \alpha(1 + i\gamma)^{1/2}]}{1 + (1 + i\gamma)^{1/2}}. \tag{3.21}$$

Two limiting cases are worthy of special note:

(I) $\alpha \to 0$, uniform n-layer, $\gamma \to \infty$.

$$\beta \doteq 2 \exp -\alpha(i\gamma)^{1/2} = 2 \exp -(1 + i)(\omega/2D)^{1/2} L. \tag{3.22}$$

(II) $\alpha \to \infty$, $q\Delta V/kT \gg 1$, $\gamma \to 0$.

$$\begin{aligned} \beta &= \exp[-i\alpha\gamma/2 - \alpha\gamma^2/8], \\ &= \exp[-i(\omega L/u) - (DL/u)/(u/\omega)^2]. \end{aligned} \tag{3.23}$$

These expressions may be interpreted as follows: In case (I), flow is by diffusion and the propagation factors k_1 and k_2 take the form $\pm(\alpha/L)(i\gamma)^{1/2}$. For this case attenuation and delay terms in the exponential are equal, and the largest negative term occurs in Z when the phase angle is $5\pi/4$ (as may be verified by differentiation.) This leads to

$$\beta = 2(-1 + i)2^{-1/2} \exp(-5\pi/4) = 0.028(-1 + i), \qquad (3.24)$$

which gives

$$Z = (1/\omega C_2)(-0.028 - i\, 1.028), \qquad (3.25)$$

the impedance of a condenser with a negative Q of 37. In order to make an oscillator by coupling this to an inductance, an inductance with a Q of more than 37 must be used. It is obviously advantageous to reduce the magnitude of the negative Q.

For case (II) in its ideal form, the ac current simply drifts through the n-layer without attenuation. This produces a phase lag of ω times the transit time L/u. If this were the only effect involved, a capacitor with a negative Q of less than unity could be produced. In addition, however, there is attenuation due to spreading by diffusion. This effect is dependent upon the ratio of the spread by diffusion $(DL/u)^{1/2}$ to the separation of planes of equal phase in the drifting hole current. This latter separation is $2\pi u/\omega$. The square of this ratio appears in the attenuation term in the second form of β.

We shall estimate the effect of the attenuation term by taking

$$\alpha\gamma/2 = 3\pi/2, \qquad (3.26)$$

so that the desired phase shift is obtained. The attenuation term is $\gamma/4$ smaller than this so that if $\gamma/4$ is considerably less than one, the attenuation in β will be small while the phase shift is correct. If we take $3\pi/2$ for the value of $\alpha\gamma/2$, then the value of γ becomes

$$\gamma = 4\omega D/u^2 = 6\pi kT/q\Delta V. \qquad (3.27)$$

Thus the approximation on which (II) is based fails unless $q\Delta V/kT > 18$, a value which implies an enormous range of concentration in the n-layer. We must, therefore, investigate the case of gradients in the n-layer by more complete algebraic procedures.

We shall denote by $-\theta_1$ the phase shift in β due to the exponential in equation (3.21). The total phase shift θ is somewhat less since the algebraic expressions give a small positive phase shift of at most about 15°, which vanishes for large and small values of γ. Similarly the attenuation of β arises chiefly from the real part of the exponential since the absolute value of the algebraic expressions lies between 2 for $\gamma = 0$ and 1 for $\gamma = \infty$.

It is instructive to express the real part of the exponent in terms of α and θ_1. This is done as follows:

$$\alpha - \alpha(1 + i\gamma)^{1/2} = -\eta - i\theta_1 . \tag{3.28}$$

This can be solved for η and $(1 + i\gamma)^{1/2}$:

$$\eta = (\alpha^2 + \theta_1^2)^{1/2} - \alpha, \tag{3.29}$$

$$(1 + i\gamma)^{1/2} = [(\alpha^2 + \theta_1^2)^{1/2} + i\theta_1]/\alpha. \tag{3.30}$$

From there it is seen that for a fixed value of θ_1, the attenuation can be greatly reduced by increasing α. Unfortunately, this requires very large changes in concentration. For example with $\theta_1 = 3\pi/2$ and $\alpha = \theta_1$, the value of η is reduced to $\eta = 0.414\ \theta_1$. However, the value of potential difference is

$$q\Delta V/kT = 3\pi, \tag{3.31}$$

giving

$$N_{d1}/N_{d2} \doteq 10^4. \tag{3.32}$$

For the case shown in Fig. 3.1, $\alpha = 1.15$ and for $\theta_1 = 5\pi/4 = 3.9$ we obtain

$$\eta = 2.95 = ln_e\, 19.5 \tag{3.33}$$

This is an improvement of about 1 factor of e in the exponential compared to having $\Delta V = 0$. The value of β is

$$\beta = 0.082 \times \exp(-i\ 218°), \tag{3.34}$$

and this leads to

$$Z = (\omega C_2)^{-1}(-0.05 - i\ 1.065). \tag{3.35}$$

Thus at the operating frequency, the diode appears to be a capacitor with a negative Q of 21.

Increasing the concentration change to a factor of 100, so that $\alpha = 2.3$, gives

$$\eta = 2.25, \tag{3.36}$$

$$\beta = 0.18 < -\ 220°, \tag{3.37}$$

$$Z = (\omega C_2)^{-1}(-0.116 - i\ 1.14), \tag{3.38}$$

$$Q = -10. \tag{3.39}$$

The calculations indicate that attenuation can be controlled to a considerable degree while maintaining the desired phase shift.

3.2. *Justification of Consequences (1), (2) and (3)*

In germanium at room temperature the product np is about 10^{27} under equilibrium conditions. At the first junction of Fig. 3.1 it is 10^{32}, implying a forward bias[3.2] of (kT/q) 2.3 × 5. In order to maintain this forward bias a flow of electrons must be furnished to N. There are several ways of accomplishing this. In the first place, the reverse bias across S_2 draws a reverse current of thermally generated electrons from P_2. This current can be controlled by controlling the lifetime and temperature in the P_2 region. Alternatively, electrons may be injected into P_2; some of these will diffuse to S_2 and arrive at N. Still another means of controlling the bias across S_1 is to make contact to N itself. Since only the dc bias need be controlled, the series resistance across N itself is unimportant; the source should be of high impedance.

The decrease in density of 10^4 across the junction in carrier concentration implies a potential difference of 9.2 (kT/q). Most of this potential difference occurs where the carrier concentration is negligible. Hence the space charge theory may be applied. Furthermore, the acceptor concentration is much higher than the donor concentration. Hence the space charge extends chiefly into the donor region and we may write[3.3]

$$\Delta V_1 = (2\pi q N_d/\kappa) W^2 \tag{3.40}$$

for the relationship between width W of the space charge region and voltage drop ΔV.

If this voltage drop has an ac component, then a charging current will be required to change W. This current is determined by the admittance

$$\omega C = \omega\kappa/4\pi W \tag{3.41}$$

of the space charge region.

At the same time injected hole and electron currents flow across the junction. The admittance associated with the hole current is approximately

$$A = (i\omega/D)^{1/2}\, \sigma_{p1}, \tag{3.42}$$

where σ_{p1} is the hole conductivity just inside the n-layer.[3.4] Actually, as discussed below, the admittance is somewhat higher.

The ratio of the admittances is

$$\frac{|A|}{\omega C} = \left[\frac{4\pi}{K}\frac{\omega\sigma_{p1}^2}{D}\cdot\frac{4\pi W^2}{\kappa\omega^2}\right]^{1/2}, \qquad = \left[\frac{4\pi\sigma_{p1}}{K\omega}\cdot\frac{q\Delta V}{kT}\cdot\frac{\sigma_{p1}}{\mu_P q N_D}\right]^{1/2}. \tag{3.43}$$

For our example this expression is much greater than 1 as may be seen as follows: The first fraction is the ratio of the dielectric decay constant to ω. This is 10^3 or more larger than ω need be. The next term is about 10 and the last term is the ratio of hole to electron density at x_1 and is about 10^{-2}. Hence the ratio of impedances is about 10:1.

We shall next consider why the expression for A for holes must be examined more closely. The admittance formula used above applies to the case of zero field to the right of the junction. The aiding field will increase the flow of holes into the n-layer and raise the admittance somewhat. Correcting for this will increase A in respect to ωC and will thus strengthen rather than weaken the argument.

Also in the expression for A, no account was taken of the transit time across the region W. If we assume a uniform field in this region for purposes of making estimates, then the solution of equation (3.20) may be applied. Since now u corresponds to drift velocity due to $9kT/q$ of voltage drop across W which is much less than L in length, it is evident that γ will be less by a large factor in this region compared to its value in N. This leads to the conclusion that phase lags will be unimportant in this region.

We have neglected the effect of electron injection into P_1. By the customary arguments for unsymmetrically doped junctions, it follows that this current is very small compared to the hole current.

This justifies consequence (1).

Consequence (2) may be justified as follows: At x_1 all the ac current $j \exp(i\omega t)$ is carried by holes. If a pure drift case occurred, the hole current might be reversed at some point in the n-layer and be $-j \exp(i\omega t)$. Under these conditions the electron current would have to be $2j \exp(i\omega t)$. Under no conditions, however, will the electron current be larger than this. This maximum possible electron current will require an electric field and this field will also affect the hole flow. Since the electron conductivity is at least 10 times larger than the hole conductivity, the hole current due to the ac field will only be about $1/10$ of j at most. Thus the hole current is only slightly affected by the ac field.

Consequence (3) follows from the fact that the reverse biased junction

S_2 has much higher impedance than S_1. S_1 has higher impedance than that for hole injection into N. The impedance for hole injection into N corresponds to the hole conductivity in the n-layer over a distance comparable with the thickness of N. However, the impedance of N itself is that due to the much larger number of electrons in it and is thus much less than the impedance of S_1. Thus it follows that the impedances across S_1 and across N are much less than across S_2. This conclusion is not affected by the modification of impedance of S_2 due to hole flow across it.

3.3. *Modifications*

The treatment presented above has been based upon the conditions (a) to (e). Some of these are advantageous from the point of view of operation but others have been introduced to simplify the treatment. Among the latter is the condition that the current across S_1 is carried chiefly by holes. If the current were chiefly capacitative at this junction, then the voltage would lag 90° behind the current. This adds a desirable phase lag in the hole injection across S_1 and thus requires less phase shift in the n-layer. By suitably adjusting the ratio of capacitative and inductive admittances, a net improvement in Q may be obtained.

4. THE TRANSIENT RESPONSE IN A UNIPOLAR STRUCTURE

In the previous section the electric field produced by the injected holes had a negligible influence on the motion of the injected holes. In effect

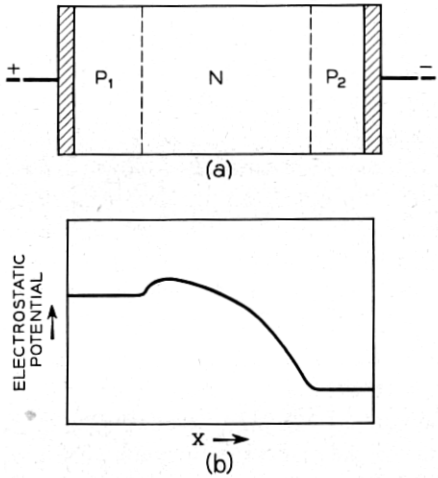

Fig. 4.1 — Space charge limited hole flow.

this was due to the bipolar nature of the mode of operation considered, the majority carriers in the region N acting to shield the minority carriers from their own space charge.

In this section we shall deal with unipolar diodes in which only one type of carrier is present in sufficient number to have a major effect. In these the influence of the space charge of the carriers upon their motion plays an important role.

Fig. 4.1 illustrates one example of the type of structure covered by the theory of this section. It is again a p-n-p structure like that considered in Section 3. However, in this case the dimensions, the donor density and the applied potential are such that the space charge "punches through" the device.[4.1] Under these circumstances a condition of space charge limited emission is set up so that holes are injected from the positive region P_1 to just such an extent that their flow is limited by their own space charge. This limitation is associated with the maximum of potential just inside N.

The potential maximum is evidently a "hook" for electrons generated thermally in P_2 and in N. Under some circumstances electrons may accumulate and form a layer in which there is no electron flow and hole flow is carried equally by diffusion and drift. Such *stagnant* regions will tend to be suppressed if P_1 is made of short lifetime material, so that electrons are siphoned out of N, or if p at the maximum is larger than p for intrinsic material and the lifetime is locally low.

We shall treat the transient response of this structure of Fig. 4.1 from the point of view of the *impulsive impedance* discussed in Section 2. Accordingly we suppose that a steady current J flows per unit area. At $t = 0$ an added pulse of current occurs carrying a total charge of δQ_i per unit area, the subscript "i" signifying initial condition. Our problem is to determine how this added charge is carried by a transient disturbance in the hole flow and what is the resultant dependence of voltage upon time; by definition the added voltage across the device is

$$v(t) = \delta Q_i D(t). \tag{4.1}$$

Since we are dealing with a planar model, we shall suppose that the initial condition at $t = 0$ corresponds to added charges δQ_i and $-\delta Q_i$ on the metal plates on the P-regions. These charges set up an added field

$$\delta E_i = \delta Q_i / K, \tag{4.2}$$

where

$$K = \kappa \epsilon_0 \tag{4.3}$$

in MKS units. The initial value $v(0)$ is then simply δE_i times the total width of the structure.

The first effect, which takes place in a negligible time in respect to the frequencies involved, is the dielectric relaxation of the field in P_1 and P_2. The added current due to δE_i leads to an exponential decay of δE in these regions with a transfer of δQ_i and $-\delta Q_i$ to the two boundaries of N. If P_1 and P_2 are thin compared to N, the resulting drop in $v(t)$ is small. In any event it can be shown by the reasoning at the end of Section 2 that this contribution to $D(t)$ adds simply the series resistance of P_1 and P_2 to the impedance.

The next effect is the transport of δQ_i on left side into N by hole flow over the potential maximum. It will be easier, however, to discuss this process after the treatment of the transient effects that occur in N itself. Consequently, we shall at this point assume that after a time, short compared with the important relaxation time in the structure, the disturbance of hole density is as shown in Fig. 4.2(a).

Fig. 4.2(a) shows added charges $+\delta Q_i$ and $-\delta Q_i$ produced by a disturbance denoted as δp in the hole density. The charge $-\delta Q_i$ on the right side is produced by an increased penetration of the space charge into P_2; it is similar to that produced by increasing reverse bias on a p-n junction.

Fig. 4.2(b) shows the corresponding disturbance in electric field. This disturbance is denoted by δE which is a function of x and t. Evidently

$$v(t) = \int_0^L \delta E(x, t)\, dx. \tag{4.4}$$

and this is the area under the δE curve.

The other parts of the figure indicate qualitatively a subsequent stage in the motion and decay of δp and δE. Our problem is to formulate mathematically this decay process. We shall treat the decay process in terms

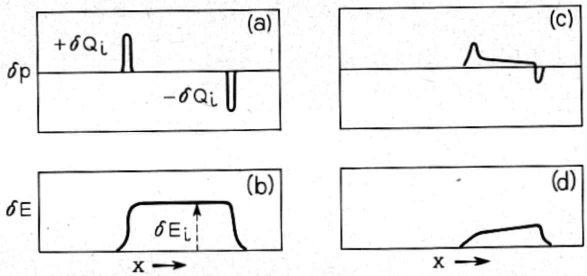

Fig. 4.2 — The initial stage and a subsequent stage of the transient.

of the effect of drift in the electric field and neglect the effects of diffusion. This procedure can be justified by the fact that as soon as a hole had reached a point where the potential has fallen by kT/q below the maximum, its flow is governed by drift rather than diffusion and the predominance of drift continues to increase towards the right.[4.2]

If drift in the field is the predominant cause of hole flow, then the equations governing the situation in N are

$$J + \delta J = (\rho + \delta\rho)(u + \delta u), \tag{4.5}$$

where the terms with δ represent the transient effects and those without represent the steady state solution, $\rho = qp$ is the charge density of the holes and u their drift velocity. The equation for the change of E with distance is

$$K(\partial/\partial x)(E + \delta E) = \rho_f + \rho + \delta\rho, \tag{4.6}$$

where ρ_f is the fixed charge density due to donors and acceptors. (We neglect any effect of traps.) The steady state equation for E is thus

$$K(dE/dx) = \rho_f + J/u. \tag{4.7}$$

In a region where ρ_f is independent of x, this equation may be reduced to quadratures by writing

$$K\, dE/(\rho_f + J/u) = dx; \tag{4.8}$$

the left side is then a known function of E through the dependence of u upon E.

It is convenient to introduce a time-like variable s which is the transit time for the dc solution. Evidently

$$ds = dx/u = K\, dE/(\rho_f u + J). \tag{4.9}$$

For the case of space charge limited current, s may be conveniently measured from the potential maximum. Even though the solution is invalid at that point, the integrals converge and the contribution from the region within kT/q of the maximum is small.

We shall assume that the equations for the steady state case have been solved and that the functional relationships are known between E, x, v and s.

The differential equation for δE may then be obtained as follows: To the left of the pulse in δp in Fig. 4.2(a), δE is zero. From equation (4.6) we have

$$K\partial E(x)/\partial x = \delta\rho. \tag{4.10}$$

Integrating this from the region where E is zero gives

$$K\delta E(x, t) = \int_0^x \delta\rho(x, t)\, dx. \tag{4.11}$$

Equation (4.11) states that the dielectric displacement at x is equal to the excess charge between the potential maximum and x. Evidently during the transient following Fig. 4.2(a), the rate of change of this extra charge is $-\delta J(x, t)$ since the dc current is flowing in at the left and an excess current δJ flows out at the right. Hence we have

$$\begin{aligned} K\partial\delta E/\partial t &= -\delta J, \\ &= -(\delta\rho u + \rho\delta u). \end{aligned} \tag{4.12}$$

For the change in drift velocity we may write

$$\delta u = (du/dE)\,\delta E = \mu^*\delta E. \tag{4.13}$$

For high electric fields u increases less rapidly than linearly with E and μ^* is less than the low-field mobility.[4.3] For very high fields μ^* is nearly zero and there are theoretical reasons for thinking that there may be a range in which μ^* is negative. We shall return to this point in the next section.

In Fig. 4.3 we show a diagrammatic representation of the transient so-

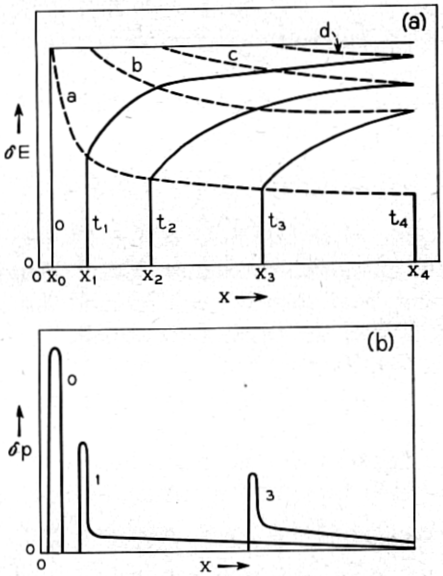

Fig. 4.3 — Graphical representation of the dependence of δE upon time.

lution. Each of the dashed lines represents the decay of δE as measured in a moving coordinate system: Thus we consider δE measured at a position $x(s_0 + t)$; this is a position that moves with the dc velocity u. This δE is evidently expressed in terms of $\delta E(x, t)$ by writing $x = x(s_0 + t)$:

$$\delta E \text{ in moving system} \equiv \delta E_m(s_0\,, t) = \delta E[x(s_0 + t), t]. \tag{4.14}$$

The differential equation for δE_m is

$$\begin{aligned} (\partial/\partial t)\, \delta E_m &= (\partial \delta E/\partial t)_x + (\partial \delta E/\partial x)_t \partial x/\partial t, \\ &= (\partial \delta E/\partial t)_x + (\partial \delta E/\partial x)_t u, \\ &= -(u\delta\rho + \rho\delta u)/K + (\delta\rho/K)u, \\ &= -(\rho\mu^*/K)\delta E = -\nu\delta E, \end{aligned} \tag{4.15}$$

where the quantity

$$\nu \equiv \rho\mu^*/m \tag{4.16}$$

is an effective dielectric relaxation constant being the *differential conductivity* $\rho\mu^*$ divided by the permittivity K.

Evidently ν is a function of position x only and may be expressed as $\nu(s)$ through the dependence of x upon s. Thus we may write

$$(\partial/\partial t)\delta E_m(s_0\,, t) = -\nu(s_0 + t)\delta E_m(s_0\,, t) \tag{4.17}$$

which has a solution

$$\delta E_m(s_0\,, t) = \delta E_m(s_0\,, 0)\ \exp\ [-g(s_0 + t) + g(s_0)], \tag{4.18}$$

where

$$g(s_0 + t) = \int_{s'}^{s_0+t} \nu(s)\, ds. \tag{4.19}$$

The lower limit s' is chosen for convenience so as to avoid singularities in $g(s)$. This integration shows that δE_m decays exponentially as the electrical field would decay in a material whose dielectric relaxation constant changed with time just as ν changes as observed on the moving plane.

Fig. 4.3 shows on the dashed lines the decay of δE_m on the moving planes. Since δE_m is zero to the left of the initial pulse in Fig. 4.2(a), it remains zero on all moving planes which follow the pulse of δQ_0. This justifies the statement made earlier. The solid curves labelled t_1, t_2 etc. show the spatial dependence of δE for times t_1, t_2, etc. after the charge δQ_1 is added.

The values of the transient voltage $v(t)$ at time t_1, for example, is the integral under the curve t_1. This curve is zero for $x < x(t_1)$ and for $x > x(t_1)$ it is

$$\delta E(x, t_1) = (\delta Q_i/K) \exp [- g(s_0 + t_1) + g(s_0)], \tag{4.20}$$

where

$$x = x(s_0 + t). \tag{4.21}$$

If the total transit time across N is S so that

$$x(S) = L, \tag{4.22}$$

then

$$v(t_1) = \int_{x(t_1)}^{L} \delta E(x, t_1)\, dx. \tag{4.23}$$

From this expression we can derive the desired formula for $D(t)$. For this purpose the integral over dx is replaced by an integral over s. At time t the range of s is evidently from t to S and $dx = u(s)\, ds$. From this we obtain:

$$\begin{aligned} D(t) &= v(t)/\delta Q_i, \\ &= (1/K) \int_s^t \exp [- g(s) + g(s - t)] v(s)\, ds. \end{aligned} \tag{4.24}$$

From Fig. 4.3 we can see that there are competing tendencies in the decay of $D(t)$ some of which tend to produce the desired convex shape discussed in Section 2 and others the concave shape. The effect of the dielectric relaxation constant is adverse and tends to produce an exponential decay. On the other hand the advance of the pulse of holes from left to right in Fig. 4.2 proceeds in an accelerated fashion with the result that the range of x over which δE is not zero decreases at an accelerated rate. If the dielectric relaxation were zero, this would result in the desired convex upwards shape.

The resultant shape of the $D(t)$ curve is thus sensitive to the exact relationship of the transit time and dielectric relaxation. This can be illustrated by giving the results of analysis for a p-n-p structure, neglecting diffusion and considering μ to be constant. The solutions of the dc equations are readily obtained for this case and have been published.[4.4] For convenience we repeat them here:

$$E = (J/\mu\rho_f)(e^{\alpha s} - 1), \tag{4.25}$$

$$x(s) = \mu J K(\mu\rho_f)^{-2} (e^{\alpha t} - \alpha t - 1), \tag{4.26}$$

$$L = x(s) = (JK/\mu\rho_f^2)(e^\beta - \beta - 1), \tag{4.27}$$

$$\beta \equiv \alpha s.$$

From these it is found that

$$lng(s) = (1 - e^{-\alpha s}). \tag{4.29}$$

This leads to

$$\begin{aligned} D(t) &= (J/\mu\rho_f^2)\,[e^\beta + e^{\alpha t}\,(\alpha t - \beta - 1)] \\ &\equiv (J/\mu\rho_f^2)\,D(\beta,\,\alpha t). \end{aligned} \tag{4.30}$$

For $t = 0$ this reduces correctly to L/K.

Figure 4.4 shows the resulting shape of the D curves with β as a parameter. Large values of β correspond to cases in which the hole charge density is small compared to ρ_f and to relatively long relaxation constants. For them the desired convex upward shape results.

Figure 4.5(a) and 4.5(b) show the real and imaginary parts of the impedance expressed in terms of $Z(\beta,\,\theta)$:

$$\begin{aligned} Z(\omega) &= \int_0^T e^{-i\omega t} D(t)\,dt \\ &= (KJ/\mu^2\rho_f^3)Z(\beta,\,\theta), \end{aligned} \tag{4.31}$$

$$\theta = \omega T. \tag{4.32}$$

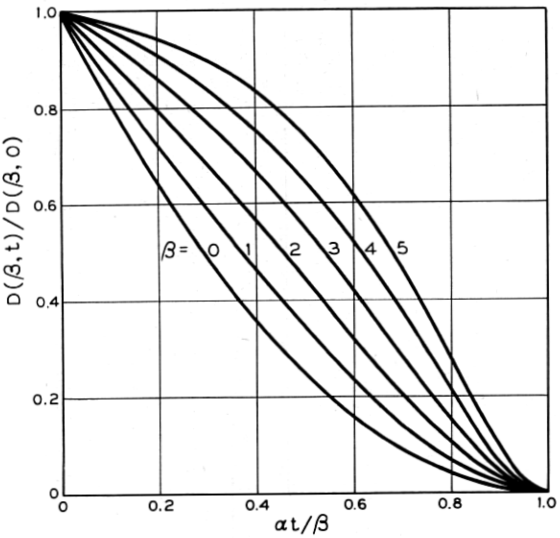

Fig. 4.4 — Impulsive impedance for various values of β in p-n-p structure.

It is seen that values of $-Q$ as small as about 10 can be obtained for $\beta \geqq 3$.

In the next section we shall consider modifications which may result from variations in μ^* and for changes in geometry.

We must return to the question of how the charge $+\delta Q_i$ passes the potential maximum. In order that the theory given above apply, it is necessary that the time required for δQ_i to enter the drift region be short compared to the transit time. At the potential maximum the charge density may be estimated by the methods previously dealt with in the theory of space charge limited emission. Initially $+\delta Q_i$ appears to the left of the maximum and the field at the maximum is δE_i. This field will

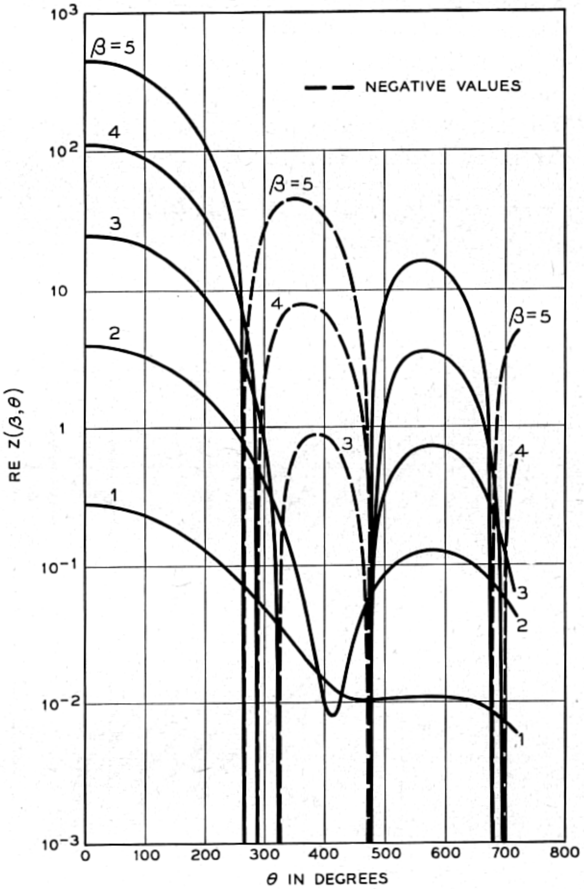

Fig. 4.5 — Impedance of a p-n-p structure. (c) Real part of impedance.

then relax with a relaxation constant of about $\mu\rho(\max)/K$ where $\rho(\max)$ is the hole charge density. Actually the relaxation may be somewhat quicker because the concentration gradient of the added holes also contributes to the flow over the maximum. Since the charge density at kT/q below the maximum is comparable to that at the maximum the entire relaxation process will proceed at about this rate. Thus a criterion for the applicability of the theory is that $K/\mu\rho(\max)$ be much less than S, the transit time or total decay time for $D(t)$.

5. MOBILITY AND GEOMETRY EFFECTS

5.1. *The Effect of a Region of Negative μ^**

In very high electric fields holes may be expected on the basis of theory to exhibit a negative value of μ^*. This theory[5.1] is founded on the idea that a hole can lose energy to phonons at a certain maximum aver-

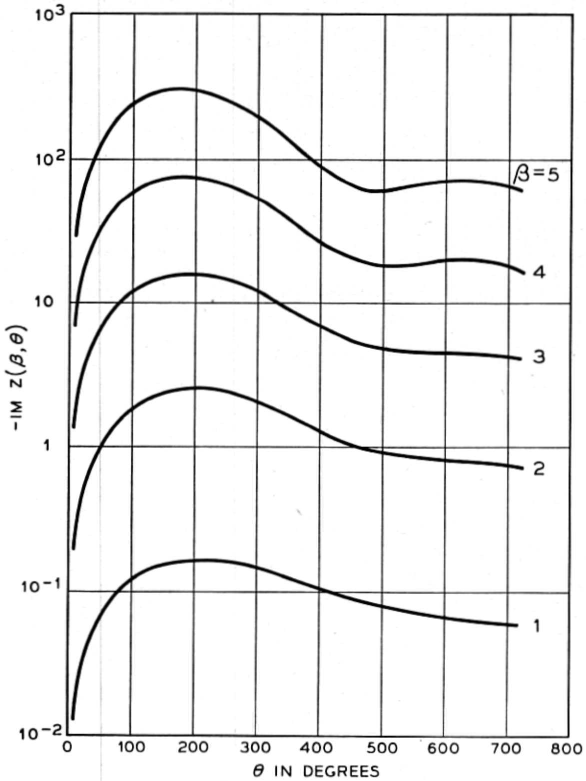

Fig. 4.5 — Impedance of a p-n-p structure. (b) Imaginary part of impedance.

age rate P_{max}. (P_{max} is the *staukonstante* of Krömer.) The holes probably achieve this rate when their energy is near the middle of the valence band. Under these conditions the power input from the electric field must be no greater than P_{max}:

$$qEu \leqq P_{max} \tag{5.1}$$

From this it follows that

$$u \leqq P_{max}/qE, \tag{5.2}$$

so that the drift velocity will decrease with increasing field at sufficiently high fields.

Furthermore, if the width of the valence band is less than the energy gap, then a hole cannot acquire enough energy to produce hole electron pairs. Thus in such a case, the negative resistance range should be reached before breakdown effects occur.

In Fig. 5.1 we illustrate the general trends of the u versus E curve, to be expected if the *staueffekt* occurs. As is indicated, the maximum drift velocity will be referred to as u_m. It occurs at a field E_m. Since we are here concerned with principles rather than details, no attempt has been made to indicate the square root range in which u is proportional to $E^{1/2}$. This range has been observed by E. J. Ryder[5.2] and shown by G. C. Dacey[5.3] to control hole flow in space charge limited hole currents in germanium and has been treated theoretically.[5.4] Dacey[5.5] has also investigated the effect of the square root law upon the $D(t)$ curves for the p-n-p structure of Section 4 and reports that the effects are so unfavorable that no negative resistance is to be expected.

The *staueffekt* opens the attractive possibility of making negative resistance devices in which the current decreases with increased dc voltage so that negative resistance will be exhibited over a wide frequency range. Unfortunately, when the boundary conditions are taken

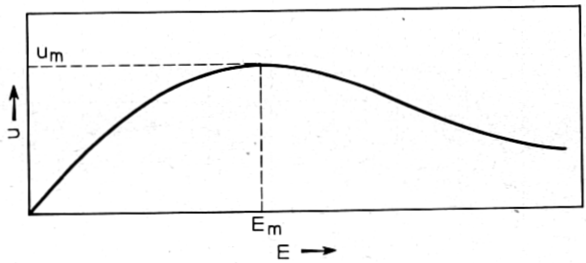

Fig. 5.1 — Qualitative representation of drift velocity versus field as affected by "staueffekt."

into account, it is found that a device in which most of the current flow occurs in a negative μ^* region does not necessarily show a dc negative resistance characteristic. On the other hand, such a structure may have a very favorable $D(t)$ characteristic.

We shall illustrate these conclusions by considering a $(p^+)p(p^+)$ structure having heavily doped ends, so that ohmic non-injecting contacts may be made to the ends. Fig. 5.2 shows the potential distribution and hole distribution for two cases of applied potential. The first case, represented in (a) and (b), corresponds to moderate fields such that the peak velocity u_m is not reached.

In the second case the voltage is so high that the average value of $E = V/L$ exceeds the critical field E_m. Under these conditions u is approximately equal to u_m over a large part of the P-region and a substantial portion of the voltage drop occurs near one end. An increase of the applied voltage occurs chiefly at this end with a small increase in current. Thus no negative dc resistance occurs.

The above conclusions are reached by considering the differential equation for the space charge again as in Section 4 neglecting diffusion and starting with $E = 0$ at the left edge of the P-region. This leads to

$$dx = K\, dE/[(J/u) - \rho_a], \tag{5.3}$$

where we have introduced

$$\rho_a = -\rho_f = -qN_a \tag{5.4}$$

for the charge density of the acceptors. Evidently if J exceeds J_m where

$$J_m = \rho_a u_m, \tag{5.5}$$

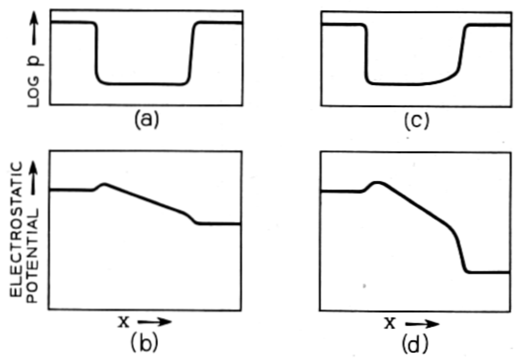

Fig. 5.2 — Hole density and potential distribution including influence of "staueffekt."

the denominator is everywhere positive and x increases monotonically with E. If J is only a little larger than J_m, there will be a large region of x in which E is nearly equal to E_m. Outside of this region, E increases much more rapidly with x. The relative scale of these distances may be estimated as follows: Suppose J is only a little larger than J_m, and consider the situation where E is about twice E_m so that u is about one half of u_m. Then

$$dx/dE \doteq K/\rho_a . \tag{5.6}$$

Under these conditions an increase of E by an additional E_m will require a distance

$$\Delta x \doteq KE_m/\rho_a . \tag{5.7}$$

If this value is much smaller than L, then the situation represented in Fig. 5.2(c) and (d) will occur. On the other hand if L is smaller than Δx, the region of space charge and high field will extend throughout most of the structure.

In any event equation (5.4) leads to positive resistance. This can be seen from the fact that increasing J always means a decrease in x for the same value of E and hence an increase in E at all values of x and thus an increase in voltage at any fixed value of x.

The above conclusion that a positive dc resistance will be exhibited by a structure like that discussed above may also be reached by considering the transient response. The theory of Section 4 may be once at be applied to this case by simply taking account of the fact that ν is negative for part of the structure and thus that δE_m increases with increasing s.

In Fig. 5.3 we illustrate a structure to which these considerations may be relatively simply applied, at least in a limiting case. It consists of four layers, the two outer being p^+ as before. Space-charge limited emission then enters the intrinsic layer which is of such a width that at its right hand boundary the electric field has a value E_3 that exceeds E_m. At this point the hole space charge is

$$\rho_3 = J/u_3 \tag{5.8}$$

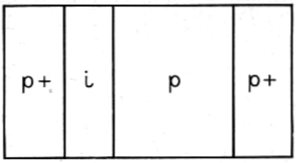

Fig. 5.3 — A structure having a region of uniform negative differential conductance.

where u_3 is $u(E_3)$. In the P-region this space charge is compensated by acceptors to produce a region of uniform field in which μ^* is negative.

If the P-region is wide compared to the I-region, then the transit time through it will also be relatively large. As a consequence δQ will be transferred quickly into the P-region. From that time on δE_m curves, like those of Fig. 4.3, will show an exponential increase with time and also with distance since for this case of constant u in the P-region, time and distance are linearly related. This will lead to a $D(t)$ of the form

$$D(t) = (u_3/K)(S - t) \exp | \mu^* \rho_3/K | t, \tag{5.9}$$

where the absolute value signs emphasize that for this case of negative μ^* there is a build-up in time. This form of D is always convex upwards and, in fact, if

$$S | \mu^* \rho_3/K | > 1, \tag{5.10}$$

it starts with a positive slope at $s = 0$ so that the transient voltage actually builds up initially with time.

Even an initially growing $D(t)$ does not give a negative resistance at low frequencies, however. As shown in Section 2, the dc resistance is simply the integral under the $D(t)$ curve and thus will still have a positive value.

5.2. *Convergent Geometry*

It is possible to obtain marked improvement of the $D(t)$ curves without the aid of the negative values of μ^*. This possibility is based upon convergent geometry. A possible case is illustrated in Fig. 5.4. In this case it is supposed that the field in the inner P-region is so large that a substantial reduction in μ^* has occurred. As a consequence, the decay of field in this region is relatively slow. Furthermore, since both the dc and

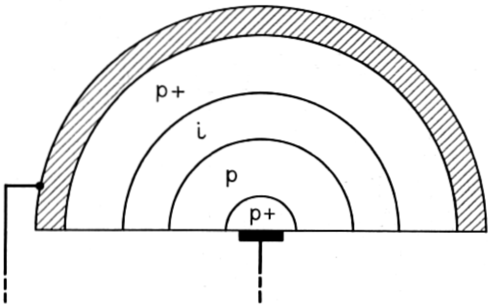

Fig. 5.4 — A convergent flow structure.

transient fields are high in this region, essentially because of the inverse square law, the principal contribution to $D(t)$ comes from this region. These two factors — relatively slow dielectric relaxation near the center and principal contribution to $D(t)$ from near the center — combine to give a $D(t)$ characteristic which holds up well until the pulse of injected holes reaches the inner region. This may result in a favorable convex upwards $D(t)$ characteristic.

ACKNOWLEDGEMENTS

The writer is indebted to a number of his colleagues for helpful discussions and to W. van Roosbroeck for the calculations for Fig. 4.4, and to R. C. Prim for Fig. 4.5.

REFERENCES

1.1. W. Shockley and W. P. Mason, J. of Appl. Phys., **25,** No. 5, p. 677, 1954.
3.1. W. Shockley, *Electrons and Holes in Semiconductors*, D. van Nostrand, N. Y., 1950, p. 312.
3.2. See Reference 3.1.
3.3. W. Shockley, B. S. T. J., **28,** p. 435, 1949, Section 2.4.
3.4. See References 3.1 or 3.3.
4.1. W. Shockley and R. C. Prim, Phys. Rev., **90,** pp. 753–758, 1953. See also G. C. Dacey, Phys. Rev., **90,** pp. 759–763, 1953, and W. Shockley, Proc. I.R.E., **40,** pp. 1289–1314, 1952.
4.2. W. Shockley and R. C. Prim, Reference 4.1.
4.3. E. J. Ryder, Phys. Rev., **90,** pp. 766–769, 1953, references.
4.4. See Shockley and Prim, Reference 4.1.
5.1. This theory of mobility in higher fields has been published by H. Krömer, Zeits f. Physik **134,** pp. 435–450, 1953. Krömer considers the theory in connection with the values of α in point contact transistors but does not explore it as a power source. The present writer derived the same result in a more primitive form in 1948 as a potential means of obtaining high frequency power. However, experimental results by E. J. Ryder did not show evidence of this effect. The effect may possibly have occurred without being recognized because of the absence of an adequate appreciation of the importance of the boundary conditions discussed in this section.
5.2. See Reference 4.3.
5.3. G. C. Dacey, Phys. Rev., **90,** pp. 759–763, 1953.
5.4. W. Shockley, B. S. T. J., **30,** pp. 990–1043, 1951.
5.5. Personal communication.