5. Internal Contact Potentials

The theory of *p-n* junctions presented above has interesting consequences when applied to the distribution of potential between two semiconductors

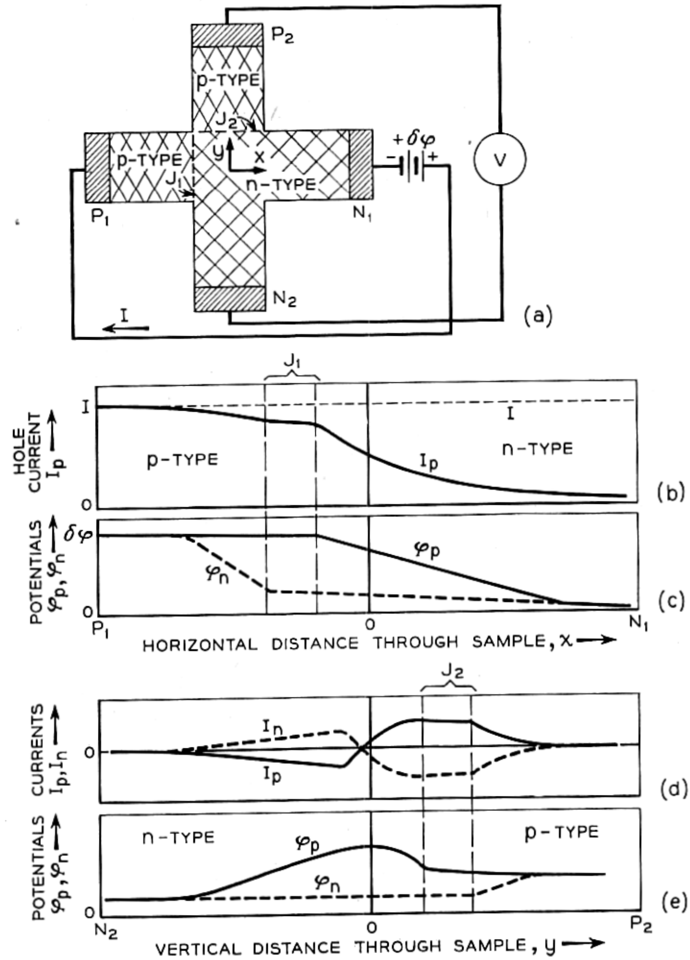

Fig. 9—Internal contact potentials showing how presence of injected holes produces a contact potential across J_2.

under conditions of hole or electron injection. In Fig. 9 we illustrate an X-shaped structure. A forward current flows across the junction P_1 and out of branch N_1. If the distance across the intersection is comparable with or small compared to the diffusion length for holes, a potential difference should be measured between P_2 and N_2. The reason for this is that holes flow easily into P_2 since the potential distribution there favors their entrance. Since, however, P_2 is open-circuited this hole flow biases J_2 in the forward direction; since J_2 is high resistance, an appreciable bias is developed before the counter current equals the inward hole flow and a steady state is reached. No similar effect occurs in the branch N_2; consequently P_2 will be found to be floating (open-circuited) at a more positive potential than N_2.

Parts (b) to (e) describe this reasoning in more complete terms. We suppose that the p-regions are more highly conducting than the n-regions so that the current across J_1, shown in (b), is mainly holes. The potentials φ_p and φ_n along the x-axis will be similar to those of Figs. 5 and 6; (c) shows this situation and indicates that the diffusion length for electrons in the p-region is less than for holes in the n-region. Along the y axis φ_p and φ_n vary as shown in (e), the reasoning being as follows: At the origin of coordinates φ_p and φ_n have the same values as for (c). The transverse hole current (d) has a small positive component at $y = 0$ since, as mentioned above, P_2 tends to absorb holes and thus increase diffusion along the plus y-axis.Since the net transverse current is zero, $I_n = -I_p$ in (d). The φ curves of (e) have been drawn to conform to the currents in (d); φ_n is nearly constant in the n-region and φ_p is nearly constant in the p-region. As concluded in connection with Figs. 5 and 6, φ_n and φ_p are also nearly constant across the transition region. These conclusions lead to the shape of φ_n and φ_p for $y > 0$ in (e). For $y < 0$, the reasoning is the same as that used in Sections 3 and 4 and we conclude that φ_n is essentially constant. Hence, a difference in the Fermi levels at P_2 and N_2 will result.

In Fig. 10 we show a structure for which we can make quantitative calculations of the variations of φ_p and φ_n. We assume for this case that the forward current from P_1 to N does not produce an appreciable voltage drop, i.e. change in ψ and φ_n, in region N. This will be a good approximation if the dimensions are suitably proportioned. We shall next solve for the steady-state distribution of p subject to the indicated boundary conditions assuming that p is a function of x only. As we have discussed in Section 4.1, when p is small compared to n in the n-region, we can write

$$p = p_n e^{q(\varphi_p - \varphi_n)/kT} \tag{5.1}$$

In keeping with the treatment in the next section of this structure as a transistor, the terminals are designated emitter, collector and base, the potentials with respect to the base being φ_ϵ and φ_c. The contact to N or the base is such that $\varphi_b = \varphi_n$ in this region. Hence, the boundary conditions at J_1 and J_2 are

$$p_1 = p_n e^{q\varphi_\epsilon/kT} \qquad x = -w \tag{5.2}$$

$$p_2 = p_n e^{q\varphi_c/kT} \qquad x = +w \tag{5.3}$$

The function $p(x)$ which satisfies these boundary conditions and the equation

$$D\frac{d^2p}{dx^2} - \frac{p - p_n}{\tau_p} = 0 \tag{5.4}$$

is

$$p(x) = p_n + \frac{p_1 + p_2 - 2p_n}{2\cosh(w/L_p)}\cosh(x/L_p) + \frac{p_2 - p_1}{2\sinh(w/L_p)}\sinh(x/L_p) \tag{5.5}$$

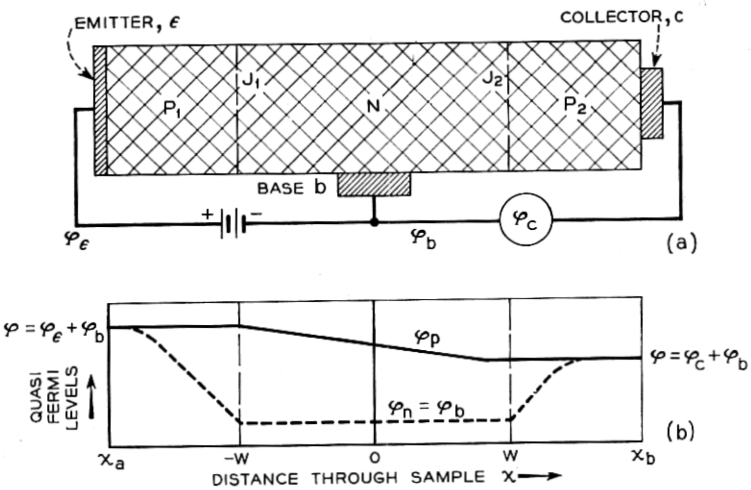

Fig. 10—Model used for calculation of internal contact potential and to illustrate p-n-p transistor.
(a) Semiconductor with two p-n junctions and ohmic metal contacts.
(b) Quasi Fermi levels showing internal contact potential between b and c.

which gives rise to a hole current across J_2 into P_2 of amount

$$\begin{aligned} I_p &= -qD\left.\frac{dp}{dx}\right|_{x=w} \\ &= \frac{qD}{2L_p}\left[(p_1 - p_2)\coth\frac{w}{L_p} + (2p_n - p_1 - p_2)\tanh\frac{w}{L_p}\right] \\ &= \frac{qD}{2L_p}\left[(p_1 - p_n)\left(\coth\frac{w}{L_p} - \tanh\frac{w}{L_p}\right)\right. \\ &\qquad\qquad \left. - (p_2 - p_n)\left(\coth\frac{w}{L_p} + \tanh\frac{w}{L_p}\right)\right] \\ &= +\frac{p_n qD}{L_p}\left[\frac{[e^{q\varphi\ /kT} - 1]}{\sinh(2w/L_p)} - \frac{(e^{q\varphi_c/kT} - 1)}{\tanh(2w/L_p)}\right] \\ &= \operatorname{csch}(2w/L_p)I_{p0}(\varphi_\epsilon) - \coth(2wL_p)I_{p0}(\varphi_c) \end{aligned} \tag{5.6}$$

where, by I_{p0} (φ), we mean the hole current which would flow in the forward direction across either J_1 or J_2 if uninfluenced by the other (i.e. the function of (4.11) or (4.18) and (4.20).) The equation shows that a fraction csch $(2w/L_p)$ of the current I_{p0} (φ_ϵ), which would be injected by φ_ϵ on P_1 in the absence of J_2, flows into P_2. The conductance of P_2 across J_2 is increased by the factor $\coth(2w/L_p)$.

The current *into* P_2 carried by electrons will be unaffected by J_1 and can be denoted by $-I_{n0}(\varphi_c)$ the minus sign resulting from the fact that currents *into* P_2 are in the reverse direction. The total current flowing *into* P_2 contains the $-I_{n0}(\varphi_c)$ and $-I_{po}(\varphi_c)$ terms and must cancel the $+I_{p0}(\varphi_e)$ term for equilibrium. Hence:

$$I_{n0}(\varphi_c) + \coth\,(2w/L_p)\,I_{p0}(\varphi_c) = \operatorname{csch}\,(2w/L_p)\,I_{p0}(\varphi_\epsilon) \tag{5.7}$$

If $p_n \gg n_p$, the I_{n0} term can be neglected compared to coth $(2w/L_p)\,I_{p0}$. Hence the value of φ_c must satisfy

$$I_{p0}(\varphi_c) = \operatorname{sech}\,(2w/L_p)\,I_{p0}(\varphi_\epsilon). \tag{5.8}$$

For $\varphi_\epsilon > kT/q$, the exponential approximation may be used for I_{p0} in both terms:

$$\varphi_c = \varphi_\epsilon - (kT/q)\ln\cosh\,(2w/L_p), \tag{5.9}$$

so that, if $(2w/L_p)$ is the order of unity, φ_c should be only about (kT/q) less than φ_ϵ. For $(2w/L_p)$ large, we get

$$\varphi_c = \varphi_\epsilon - (kT/q)\,(2w/L_p) \tag{5.10}$$

corresponding to the linear drop of φ_p, discussed in connection with equation (4.9), across the distance $2w$.

When φ_ϵ is negative, so that we have to deal with reverse current, φ_c will not decrease indefinitely but will reach a minimum value given by

$$[\exp q\varphi_c/kT] - 1 = -\operatorname{sech}\,(2w/L_p) \tag{5.11}$$

and corresponding to saturation reverse current across J_1, so that

$$\varphi_c = -(kT/q)\ln\,[1 + (1/2)\operatorname{csch}^2(w/L_p)]. \tag{5.12}$$

The floating potentials of p-type contacts to n-type material into which holes have been injected (or n-type contacts to p-type material with injected electrons) are reminiscent of probes in gas discharges which tend to become charged negative in respect to the space around them because they catch electrons more easily than positive ions. The situation may also be compared with that producing thermal e.m.f.'s; in fact a "concentration temperature" of the semiconductor with injected holes can be defined by finding the temperature for which $np = n_i^2(T)$. We conclude that, in the absence of thermal equilibrium, different potentials depending on the nature of the contact are, in general, the rule rather than the exception.

The bias developed on P_2 or c will change its conductance. If we suppose that φ_ϵ and φ_b are held constant, then the current flowing into c is obtained by the same reasoning that led to (5.7) and is

$$I_c(\varphi_c, \varphi_\epsilon) = I_{nC}(\varphi_c) + \coth \frac{2w}{L_p} I_{p0}(\varphi_c) - \operatorname{csch} \frac{2w}{L_p} I_{p0}(\varphi_\epsilon). \quad (5.13)$$

For an infinitesimal change in φ_c from the value which makes $I_c(\varphi_c, \varphi_\epsilon)$ vanish, the admittance to c is readily found from (4.18) and (4.19) to be

$$\begin{aligned} \left(\frac{\partial I_c}{\partial \varphi_c}\right)_{\varphi_\epsilon} &= I'_{n0}(\varphi_c) + \coth \frac{2w}{L_p} I'_{p0}(\varphi_c) \\ &= \left[G_{n0} + \coth \frac{2w}{L_p} G_{p0}\right] e^{q\varphi_c/kT} \end{aligned} \quad (5.14)$$

which shows that pronounced variations in admittance should be associated with variations in hole density in N in Fig. 10.[15]

6. *p-n-p* Transistors

The structure shown in Fig. 10 is a transistor with power gain provided the distance w is not too great. As a first approximation, we shall neglect the drop due to currents in the N region. If we use P_2 as the collector and call the collector current, I_c, positive when it flows into P_2 from outside, we shall have from (5.13)

$$I_c = -\operatorname{csch} \frac{2w}{L_p} I_{p0}(\varphi_\epsilon) + \coth \frac{2w}{L_p} I_{p0}(\varphi_c) + I_{n0}(\varphi_c). \quad (6.1)$$

The emitter current is similarly

$$I_\epsilon = \coth \frac{2w}{L_p} I_{p0}(\varphi_\epsilon) - \operatorname{csch} \frac{2w}{L_p} I_{p0}(\varphi_c) + I_{n0}(\varphi_\epsilon). \quad (6.2)$$

If $p_n \gg n_p$, then the I_{n0} terms can be neglected. However, the base current will not vanish but will be

$$\begin{aligned} I_b = -I_\epsilon - I_c &= \left[\operatorname{csch} \frac{2w}{L_p} - \coth \frac{2w}{L_p}\right] [I_{p0}(\varphi_\epsilon) + I_{pC}(\varphi_c)] \\ &= \frac{2 \sinh^2 w/L_p}{\sinh 2w/L_p} [I_{p0}(\varphi_\epsilon) + I_{p0}(\varphi_c)]. \end{aligned} \quad (6.3)$$

[15] The variations in admittance discussed in connection with metal point contacts in an accompanying paper in this issue (W. Shockley, G. L. Pearson and J. R. Haynes, *Bell Sys. Tech. Jl.*, July, 1949), arise from this cause; however, the nature of the contact is not as simple as here.

For w/L_p large, the junctions do not interact and the hyperbolic coefficient becomes unity and $I_b = -[I_{p0}(\varphi_\epsilon) + I_{p0}(\varphi_c)]$.

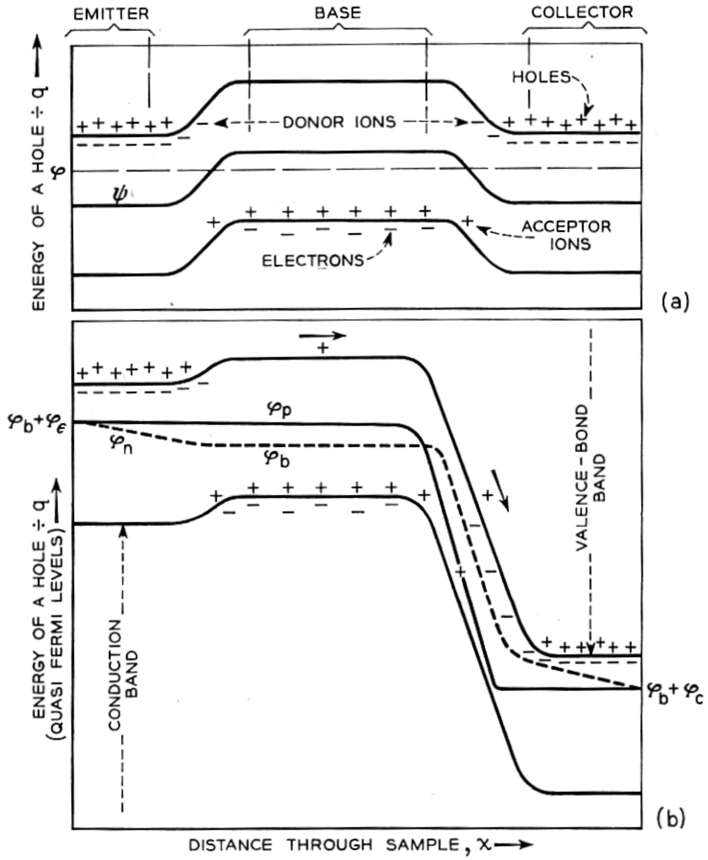

Fig. 11—*p-n-p* transistor.
(a) Thermal equilibrium.
(b) Operating condition.

If φ_c is several volts negative, so that $I_{p0}(\varphi_c)$ has its saturation value I_{ps} (see (4.11) and (4.20)), then the ratio $-\delta I_c/\delta I_\epsilon \equiv \alpha$ has the value

$$\alpha = -\frac{\delta I_c}{\delta I_\epsilon} = \frac{\operatorname{csch}\dfrac{2w}{L_p}}{\coth\dfrac{2w}{L_p}} = \operatorname{sech}\frac{2w}{L_p}. \tag{6.4}$$

For $(2w/L_p) = 0.5, 1, 2$ respectively, $\alpha = 0.89, 0.65, 0.27$. Since the output impedance R_{22} will be very high when φ_c is in the reverse direction, and the input impedance will be low, the power gain formula[16] $\alpha^2 R_{22}/R_{11}$ will yield power gain even when α is less than unity.

In certain ways the structure of Fig. 10 resembles a vacuum tube. In Fig. 11, we show the energy band diagram, with energies of holes plotted upwards so as to be in accord with the convention for voltages. (a) shows the thermal equilibrium distribution and (b) the distribution under operating conditions. It is seen that the potential hill, which holes must climb in reaching the collector, has been reduced by φ_ϵ. The *n*-region represents in a sense the grid region in a vacuum tube, in which the potential and hence plate current, is varied by the charge on the grid wires. Here the potential in the *n*-region is varied by the voltage applied between base and emitter. In both cases one current is controlled by another. In the vacuum tube the current which charges the grid wires controls the space current. Because the grid is negative to the cathode, the electrons involved in the space current are kept away from the grid while at the same time the electrons in the grid are kept out of the space by the work function of the grid (provided that the grid does not become overheated.) In Fig. 11, the electrons flowing into the base control the hole current from emitter to collector. In this case the controlled and controlling currents flow in the same space but in different directions because of the opposite signs of their charges.

As this discussion suggests, it may be advantageous to operate the *p-n-p* transistor like a grounded cathode vacuum tube, with the emitter grounded and the input applied to the base.

The *p-n-p* transistor has the interesting feature of being calculable to a high degree. One can consider such questions as the relative ratios of width to length of the *n*-region and the effect of altering impurity contents and scaling the structure to operate in different frequency ranges. However, we shall not pursue these questions of possible applications further here.

Acknowledgment

The writer is indebted to a number of his colleagues for stimulating discussions and encouragement, in particular to H. R. Moore, G. L. Pearson and M. Sparks, whose experimental work, to be described in later publications, suggested development of the theory along the lines presented above. He is also indebted to J. Bardeen, P. Debye, G. Wannier and W. van Roosbroeck for theoretical comments and suggestions and especially to the last and to Mrs. G. V. Smith for valuable assistance in preparing the manuscript.

[16] Physical Principles Involved in Transistor Action, J. Bardeen and W. H. Brattain, *Phys. Rev.* 75, 1208 (1949).

APPENDIX I

A Theorem on Junction Resistance

We shall here prove that the junction resistance is never less than the value obtained by integrating the local resistivity $1/q\mu(p + bn)$. This is accomplished by analyzing the following equation which we shall discuss before giving the derivation:

$$I\delta\varphi = \frac{1}{q\mu}\int_{x_a}^{x_b}\left(\frac{I_p^2}{p} + \frac{I_n^2}{bn}\right)dx + qg\int_{x_a}^{x_b}(\varphi_p - \varphi_n)(e^{q(\varphi_p-\varphi_n)/kT} - 1)\,dx,$$

the meaning of the symbols being that shown in Fig. 5. This expression is valid even if large disturbances in p and n from their equilibrium values occur. The second integral is positive since the integrand is never negative. It may be very large if $\varphi_p - \varphi_n \gg kT/q$ in some regions. If, in the first integral, we consider that I_p and I_n may be varied subject to the restraint $I_p + I_n = I$, we may readily prove that the first integrand takes on a minimum value when

$$I_p = \frac{pI}{p + bn} \quad \text{and} \quad I_n = \frac{bnI}{p + bn}.$$

For this minimum condition, the first integral becomes

$$I^2\int_{x_a}^{x_b} dx/q\mu(p + bn) = I^2 R_0$$

where R_0 is simply the integrated local resistivity. If I does divide in this way, the second integral is zero, a result which we can see as follows:

$$I_p = -q\mu p\, d\varphi_p/dx$$

$$I_n = -q\mu bn\, d\varphi_n/dx$$

$$\frac{d\varphi_p/dx}{d\varphi_n/dx} = \frac{I_p/p}{I_n/bn}$$

Hence, if the current divides in the ratio of p to bn, then $d\varphi_p = d\varphi_n$ and, since $\varphi_p = \varphi_n$ at x_a, $\varphi_p = \varphi_n$ everywhere and the second integral vanishes.

In general, of course, the conditions governing recombination prevent current division in the ratio $p{:}bn$ and then $\delta\varphi/I > R_0$.

The equation discussed above is derived as follows: We suppose that

$$\varphi_p(x_a) = \varphi_n(x_a) = \varphi_a$$

$$\varphi_p(x_b) = \varphi_n(x_b) = \varphi_b$$

Then

$$I\delta\varphi = -(I_p\varphi_p + I_n\varphi_n)\Big|_{x_a}^{x_b}$$

$$= -\int_{x_a}^{x_b} \frac{d}{dx}(I_p\varphi_p + I_n\varphi_n)\,dx$$

$$= -\int_{x_a}^{x_b}\left(\frac{dI_p}{dx}\varphi_p + \frac{dI_n}{dx}\varphi_n\right)dx - \int_{x_a}^{x_b}\left(I_p\frac{d\varphi_p}{dx} + I_n\frac{d\varphi_n}{dx}\right)dx.$$

Since

$$\frac{dI_p}{dx} = -\frac{dI_n}{dx} = qg(1 - e^{q(\varphi_p-\varphi_n)/kT})$$

and

$$\frac{d\varphi_p}{dx} = -I_p/q\mu p, \qquad \frac{d\varphi_n}{dx} = -I_n/q\mu bn$$

these two integrals are readily transformed into the ones previously discussed.

APPENDIX II

Admittance in a Retarding Field

We shall here derive the admittance equation for holes diffusing into a retarding potential $\psi = kTx/qL_r$ in which the potential increases by kT in each distance L_r. The differential equation for the a-c. component of p is

$$i\omega p = -\frac{p}{\tau_p} - \frac{\partial}{\partial x}\left[-D\frac{\partial p}{\partial x} - \mu p\frac{\partial\psi}{\partial x}\right].$$

This equation may be solved by letting $p = p_1\exp(i\omega t - \gamma x)$ as may be seen by rewriting the equation and substituting this expression for p:

$$D\left[\frac{\partial^2 p}{\partial x^2} + \frac{1}{L_r}\frac{\partial p}{\partial x}\right] - \frac{1}{\tau_p}(1 + i\omega\tau_p)p$$

$$= -\gamma D\left[-\gamma + \frac{1}{L_r}\right]p - \frac{1}{\tau_p}(1 + i\omega\tau_p)p = 0$$

leading to

$$\gamma = \frac{1 + [1 + (2L_r/L_p)^2(1 + i\omega\tau_p)]^{1/2}}{2L_r}.$$

The corresponding current evaluated at $x = 0$ where $p = p_1 \exp(i\omega t) = (p_n q v_1/kT)\exp(i\omega t)$ is given by

$$\begin{aligned} I &= -q\left[D\frac{\partial p}{\partial x} + \mu p\frac{\partial \psi}{\partial x}\right] \\ &= -qD\left[-\gamma + \frac{1}{L_r}\right]p \\ &= \frac{q(1 + i\omega\tau_p)p}{\gamma\tau_p} \\ &= \frac{p_n q^2(1 + i\omega t)}{kT\tau_p} \cdot \frac{2L_r}{1 + [1 + (2L_r/L_p)^2(1 + i\omega\tau_p)]^{1/2}} \cdot v_1 e^{i\omega t} \\ &= \frac{q\mu p_n 2L_r}{L_p^2} \cdot \frac{(1 + i\omega\tau_p)}{1 + [1 + (2L_r/L_p)^2(1 + i\omega\tau_p)]^{1/2}} \cdot v_1 e^{i\omega t} \\ &= A_p v_1 e^{i\omega t}. \end{aligned}$$

This is equivalent to (4.32) in Section 4.

APPENDIX III

Admittance for Two Layers

We shall here treat a case in which there is a thin layer on the n-side of the transition region in which recombination occurs much more readily than deeper in the n-layer. The case of an infinitely thin plane, discussed in Section 4, is a limiting case of this model. We shall suppose that the layer extends from $x = -c$ to $x = 0$ while $x > 0$ corresponds to the n-region. We shall suppose that the potential in the layer is uniform with value ψ_1 whereas in the n-region it has value ψ_2. The lifetimes of holes will be taken τ_1 and τ_2 in the two layers. The solutions for p_1 and p_2 are evidently

$$p_1 = p_{10} + (A\,e^{-\alpha x} + B\,e^{+\alpha x})\,e^{i\omega t} \qquad x < 0$$

$$p_2 = p_{20} + C\,e^{-\beta x + i\omega t} \qquad x > 0$$

where

$$\alpha = (1 + i\omega\tau_1)^{1/2}/\sqrt{D\tau_1} \equiv (1 + i\omega\tau_1)^{1/2}/L_1$$

$$\beta = (1 + i\omega\tau_2)^{1/2}/\sqrt{D\tau_2} \equiv (1 + i\omega\tau_2)^{1/2}/L_2.$$

The boundary condition for continuity of φ_p, required to avoid singularity in $\partial\varphi_p/\partial x$, is

$$p_2 e^{q\psi_2/kT} = p_1 e^{q\psi_1/kT}$$

and, for continuity of hole current, is $\partial p_1/\partial x = \partial p_2/\partial x$. Expressing these in terms of A, B, C, α and β for the a-c. components yields:

$$A + B = Ce^{q(\psi_1 - \psi_2)/kT} \equiv CF$$

$$\alpha(A - B) = \beta C$$

so that

$$A = (F + \beta/\alpha)C/2.$$

$$B = (F - \beta/\alpha)C/2.$$

Hence the ratio $-[\partial p/\partial x]/p$ at $x = -c$ is

$$-\frac{\partial \ln p}{\partial x} = \frac{\alpha(A e^{+\alpha c} - B e^{-\alpha c})}{(A e^{+\alpha c} + B e^{-\alpha c})} = \frac{\alpha(F\alpha \sinh \alpha c + \beta \cosh \alpha c)}{F\alpha \cosh \alpha c + \beta \sinh \alpha c}.$$

Since at $x = -c$, the a-c. component of p_1 is $(qv_1/kT)p_{10}e^{i\omega t}$, the admittance is

$$A_p = \frac{-qD\partial p/\partial x}{v_1 e^{i\omega t}} = (q^2 Dp_{10}/kT)(-\partial \ln p/\partial x)$$

$$= (q\mu p_{10}/L_1)(1 + i\omega\tau_1)^{1/2} \frac{F\alpha \sinh \alpha c + \beta \cosh \alpha c}{F\alpha \cosh \alpha c + \beta \sinh \alpha c}.$$

For $c \to 0$, this transforms into

$$(q\mu p_{10}/L_1)\ (1 + i\omega\tau_1)^{1/2}\beta/F\alpha = (q\mu(p_{10}/F)/L_2)(1 + i\omega\tau_2)^{1/2}$$

which agrees with Section 4, since p_{10}/F then corresponds to p_n.

If c/L_1 and F are not large, an appreciable amount of recombination takes place for $x > 0$ for low frequencies. Dispersive effects will then occur corresponding to τ_2. The a-c. will not penetrate to $x = 0$, however, if $c(\omega/D)^{1/2} \gg 1$ and the dispersive effects will then be determined by τ_1.

The frequency-dependent part of the admittance,

$$(1 + i\omega\tau_1)\ \frac{F\alpha \sinh \alpha c + \beta \cosh \alpha c}{F\alpha \cosh \alpha c + \beta \sinh \alpha c},$$

has been computed and is shown in Fig. 7 for $\tau_p = \tau_2$, $F = 1$, $\tau_1 = \tau_2/9$ and $c/L_1 = \frac{1}{3}$. For these values about half the hole current reaches $x = 0$ for low frequencies. As the time constant for diffusion through the layer is $\tau_p/81$, as discussed in Section 4.6, the layer will act as a largely frequency-independent admittance well above the point for $\omega\tau_p = 1$. This is reflected in the behavior of the curves of Fig. 7 and, for frequencies in the $\sqrt{\omega t}$ range, it is seen that G is larger than S by about 50% of the low-frequency value of G; this split of $G + iS$ into $(\frac{1}{2})G_0$ plus approximately $(\frac{1}{2})G_0\,(1 + i\omega\tau_p)^{1/2}$ corresponds to the fact that about half the holes are absorbed in layer 1 for the assumed conditions.