

A MODERN APPROACH TO SEMICONDUCTOR AND VACUUM DEVICE THEORY

By R. D. MIDDLEBROOK, M.A., M.S., Ph.D.

(The paper was presented at the INTERNATIONAL CONVENTION ON TRANSISTORS AND ASSOCIATED SEMICONDUCTOR DEVICES, 22nd May, 1959.
The written version was received 28th July, 1959.)

SUMMARY

An integrated approach to the understanding of charge-controlled electronic devices is presented. Although only vacuum triodes and diffusion-type transistors are discussed in detail, the methods suggested are also applicable to gas-filled and multi-electrode vacuum structures, to surface-barrier and to drift-type transistors, and to space-charge-limited solid-state devices. The treatment is tutorial in nature, and begins with the development of general equations of current flow applicable in any medium. The principles of charge-controlled devices are then summarized, and a general functional relationship between the total charge in transit and the transit time is developed. These results are then applied in turn to vacuum and semiconductor diodes and triodes to derive in a remarkably simple and consistent manner the salient features of their operation. 'Ideal' vacuum triode and transistor structures are first discussed, and the voltage and current amplification factors are then introduced as arbitrary parameters to account for practical departures from ideality. Specific results obtained are the d.c. characteristics and incremental equivalent circuits for each device. The model established for the transistor is identical with the hybrid- π circuit due to Giacoletto, and both low- and high-level injection conditions are included. Finally, it is suggested that the transistor collector saturation current with open base is a more fundamental quantity than that with open emitter, and the temperature dependence of the base-emitter voltage is shown to be linear at any injection level. Throughout, emphasis is on the principles involved and on the method of approach, and a particular effort is made to present the development of the vacuum and the semiconductor devices in a completely analogous manner.

LIST OF PRINCIPAL SYMBOLS

For simplicity in mathematics, the unit-area approach is used throughout; the few departures from this will be obvious.

C_1, C_2 = Input and output intrinsic capacitances.

C_{eg}, C_{ep}, C_{ec} = Capacitances from equivalent grid plane to grid, anode and cathode (Fig. 6).

$C_t(C_{tc}, C_{te})$ = Transition-region capacitance (at collector and emitter).

$D(D_n, D_p)$ = Diffusion constant (of electrons and holes).

E = Electric field.

G_1, G_2 = Input and output intrinsic conductances.

I_1, I_2 = Input and output current densities.

I_s = Saturation current density [eqn. (91), metal-semiconductor; eqn. (118), p - n junction].

N_a = Acceptor density.

N_d = Donor density.

Q = Total charge in transit.

T = Absolute temperature.

V = Potential.

V_1, V_2 = Input and output voltages.

V_a = Vacuum-diode anode voltage.

V_b = (i) Semiconductor diode voltage (Sections 5 and 6.1). (ii) Base-emitter voltage (Section 6.2).

V_c = Collector voltage.

V_e = Equivalent grid-plane voltage.

V_g = Grid voltage.

V_j = Voltage across transition region.

V_p = (i) Anode voltage (Section 4). (ii) Voltage across charge-neutral p -region (Section 6).

V_s = Diffusion potential (Fig. 8).

W_b = Barrier height (Fig. 8).

W_g = Energy gap between valence and conduction bands.

W_f = Fermi level.

$W_t = kT/e$.

ϕ_m = Work function of a metal.

ϕ_s = Work function of a semiconductor.

$Z = n_0/p_p$, injection-level parameter.

d_1, d_2 = Distances from grid plane to cathode and plate.

e = Magnitude of electronic charge.

g_m = Transconductance.

i = Current density.

i_n = Electron current density.

k = Boltzmann's constant.

m = Mass of electron.

n = (i) Exponent in relation $\tau_t \propto 1/Q^n$. (ii) Electron density.

n_0 = Injected electron density.

n_i = Intrinsic carrier density.

n_n, n_p = Equilibrium electron density in n -region and p -region.

p = Hole density.

p_0 = Hole density at edge of transition region [eqn. (102)].

p_n, p_p = Equilibrium hole density in n -region and p -region.

v = Charge velocity.

w = (i) Anode-cathode distance in vacuum diode (Section 4). (ii) Thickness of charge-neutral p -region (Section 6).

w_s = Transition region thickness in equilibrium (Figs. 8 and 10).

x_t = Transition-region thickness.

β = Ratio of incremental collector and base currents.

δ = Geometry-dependent parameter defined in eqn. (78).

$\epsilon(\epsilon_0)$ = Permittivity (of free space).

$\mu \equiv C_{eg}/C_{ep} = C_1/C_2$ = Ratio of incremental anode and grid voltages at constant current.

τ_t = Average transit time.

ρ = Charge density.

ρ_a = Fixed negative charge density.

ρ_d = Fixed positive charge density.

ρ_e = Net charge density.

ρ_n = Mobile negative charge density.

ρ_p = Mobile positive charge density.

$\mu(\mu_n, \mu_p)$ = Mobility (of electrons and holes).

Dr. Middlebrook is Associate Professor of Electrical Engineering at the California Institute of Technology, Pasadena, California.

(1) INTRODUCTION

Whenever a new field of endeavour is opened up, much effort is expended in investigating all possible avenues of advance in the understanding and application of the new ideas. Inevitably some of these avenues are later recognized to be more fundamental or important than others: only the cumulative experience of many workers can lead to a realization of the proper perspective into which the many separate items of knowledge in the field should be placed.

Such a crystallization of the salient features in the theory and application of the ordinary vacuum tube has long since been attained. This happy condition is at present evolving in the theory and application of transistors, although the process is far from complete, since new semiconductor devices continue to appear at a rapid rate. Upon contemplation of the basic theories of the vacuum tube and the transistor, it soon becomes apparent that a further step in the evolution of each is desirable—that of combining the two theories into one, in which the tube and the transistor are examples of a general active device.

Although many workers have suggested similarities between tubes and transistors, the way to a really integrated treatment of both devices has been suggested only recently by Johnson and Rose¹ in the United States and by Sparkes and Beaufoy² in England, who have pointed out that the tube and the transistor are both fundamentally charge-controlled devices, and not, as had previously been believed, voltage-controlled and current-controlled devices respectively.

The paper represents an attempt to follow up the basic charge-control approach, to integrate the treatment of current flow in vacuo and in a solid, and to derive in a remarkably simple manner the salient features of vacuum and semiconductor diodes and triodes. In Section 2 the basic equations of current flow are established, applicable to charge motion in vacuo, in solids, and in electrolytes. The Einstein relation between the mobility and the diffusion constant of a carrier follows directly from a discussion of the mechanisms of drift and diffusion flow. In Section 3 some results of Johnson and Rose on the basic properties of charge-controlled devices are reviewed, and it is shown that the transit time of charged carriers across the active region is proportional to the *n*th power of the total charge in transit, where *n* = 0 when the current is diffusion limited, 0.5 when it is space-charge limited in vacuo, and 1 when it is space-charge limited in a solid. A simple incremental equivalent circuit is presented which is independent of the type of charge-controlled device.

The basic current-flow equations are applied to vacuum diodes and triodes in Section 4, to obtain the fundamental relationship between external voltages and currents in terms of total charge and transit time. The properties of the practical vacuum triode are obtained very simply from those of an 'ideal' triode (whose grid draws no current and exerts sole control over the anode current) by introducing the voltage amplification factor as an arbitrary geometry-dependent parameter. Application of the results of Section 2 leads to equations for the element values in incremental models of the vacuum diode and triode.

In Section 5 some basic properties of metal-semiconductor junctions are considered in order to introduce the important assumption of a sharp discontinuity between charge-depletion and charge-neutral regions in a semiconductor. The voltage dependence of the thickness of the charge-depletion region at a junction is derived, and the rectifier characteristic and transition-region capacitance are discussed.

The basic current-flow equations are applied to semiconductor diodes and triodes in Section 6. The procedure of applying these equations to the charge-depletion region and to the charge-neutral region separately, and then matching boundary condi-

tions, is emphasized. General but simple results applicable to both low- and high-level injection are presented. The properties of the practical semiconductor triode (transistor) are obtained very simply from those of an 'ideal' triode (whose base draws no current and exerts sole control over the collector current) by introducing the current amplification factor as an arbitrary geometry-dependent parameter. A feature of the treatment is that the transistor is discussed throughout in terms of the common-emitter configuration, so avoiding the artificial and unrealistic technique of first deriving all the parameters in the common-base connection and then transforming them for the common-emitter one. Application of the results of Section 2 leads to an incremental equivalent model identical with the hybrid- π model introduced by Giacoletto.³ Equations for the model element values are given for both low- and high-level injection conditions. Finally, the temperature dependences of the collector saturation current and base to emitter voltage are discussed, and it is suggested that the collector current with open base is a more fundamental quantity than that with open emitter.

Although only a few specific devices are discussed in the paper, the general approach is applicable to all types of charge-controlled device, such as gas tubes and space-charge-limited solid-state structures. Some comparisons between gaseous and semiconductor devices have already been drawn by Webster,⁴ and the analogy between space-charge-limited current flow in vacuo and in a semiconductor was first discussed by Shockley and Prim.⁵ It is felt that the approach here described is capable of even further generalization in order to integrate the theory of a large class of electronic devices into a complete whole.

(2) BASIC EQUATIONS OF CURRENT FLOW

In order to cause a current to flow, a potential gradient must exist. To determine the current density, *i*, at any point, the charge density, ρ , and charge velocity, *v*, must be known. There are thus, in general, four unknowns involved in the solution for the current in a given physical device, namely *i*, ρ , *v*, and *V*, some or all of which may be functions of positions and of time, and so four equations relating these variables are required (magnetic fields are not considered). The simultaneous solution of these equations will be subject to boundary conditions imposed by the particular physical device.

In a given structure the sources of charge density may be both mobile and fixed charges of both signs. Thus distinction must be made between the net charge density, ρ_e , which is the algebraic sum of the charge densities of all sources, and the charge density ρ due to one source. With this in mind, three of the four required equations between the four variables may be stated as follows:

$$\text{Maxwell's equation: } i = \rho v + \epsilon \frac{\partial E}{\partial t} \dots \dots \dots (1)$$

$$\text{Poisson's equation: } \nabla^2 V = - \rho_e / \epsilon \dots \dots \dots (2)$$

Continuity equation:

$$\frac{\partial \rho}{\partial t} = e(g - r) - \nabla \cdot (\rho v) \dots \dots \dots (3)$$

together with the subsidiary relation

$$\rho_e = \rho_p + \rho_n + \rho_d + \rho_a \dots \dots \dots (4)$$

The ρ which appears in the first and third equations is to be taken as ρ_p or ρ_n depending on whether current flow due to positive or to negative charges is to be considered, and *g* and *r* are respectively the rate of appearance and disappearance of

the charge density ρ_p or ρ_n as appropriate. In specific physical circumstances, eqn. (4) may be expressed in terms of more familiar quantities as follows. For electrons in a vacuum, ρ_p and ρ_a are each zero and $\rho_n = -en$. For electrons and positive ions in a vacuum or for negative and positive ions in an electrolyte, ρ_d and ρ_a are each zero, and $\rho_p = ep$ and $\rho_n = -en$. For electrons and holes in a semiconductor, $\rho_d = eN_d$ and $\rho_a = -eN_a$, and $\rho_p = ep$ and $\rho_n = -en$.

The fourth required equation between the variables i , ρ , v and V involves the detailed mechanism by which the velocity is imparted to the charge carrier. Since usually a large number of carriers will be present, in random thermal motion, it becomes necessary to consider two limiting cases: one in which the dimensions of the device of interest are short and one in which they are long compared with the mean free path between thermal collisions. In the first case each carrier can be considered individually, and in the second only the average motion of large numbers of carriers can be usefully discussed. Fortunately, one or the other of the limiting cases is well approximated in practical structures.

When the device dimensions are small compared with the mean free path, the velocity attained by a mobile carrier (assumed negatively charged) is easily found from the basic relations

$$f = m \frac{dv}{dt} = mv \frac{dv}{dr} \quad \dots \dots \dots (5)$$

$$f = e \frac{dV}{dr} \quad \dots \dots \dots (6)$$

where f is the force on the carrier due to the potential gradient and r is the direction of the potential gradient. From eqns. (5) and (6) the velocity attained by a carrier starting from rest at zero potential is

$$v = \sqrt{\frac{2e}{m} V^{1/2}} \quad \dots \dots \dots (7)$$

This is thus the fourth relation between the variables, for the given condition. It may also be noted that under these conditions the acceleration is constant and the velocity increases linearly with time in a constant potential gradient. Current resulting from charge velocity governed by these conditions may be called a free-acceleration current.

Statistical methods must be used under conditions in which the device dimensions are large compared with the mean free path, when each carrier suffers many collisions in transit through the device. In the presence of a potential gradient a charge carrier will be accelerated in accordance with eqn. (5), but will suffer a collision after travelling, on the average, one mean free path. If the energy gained from the field during the acceleration is small compared with the thermal energy, the velocity after the collision will be random in direction. On the average, therefore, a carrier starts from rest during each interval between collisions and attains (superimposed on the random thermal velocity) an average drift velocity over many such intervals which is proportional to the potential gradient. This assumption breaks down at high fields, since the condition mentioned above is no longer valid.

A carrier can acquire an average velocity superimposed on its random thermal velocity as a result of either a potential gradient or a density gradient. The argument in justification of this statement proceeds as follows. As a result of thermal collisions, either between themselves or with a fixed crystal lattice, carriers tend to diffuse away from regions of higher density. This motion may be imagined as being due to a 'density-gradient force', and if the carriers are charged, such motion constitutes a diffusion

current. The average density-gradient force on each carrier can be considered to impart an acceleration to the carrier just as does a potential-gradient force. The fundamental similarity of drift current in a potential gradient and diffusion current in a density gradient may be made more apparent by the following derivation of quantitative expressions for the two types of carrier flow. The philosophy of this argument has been anticipated by Parker in connection with electron and ion drift and diffusion in plasmas.⁶

Consider negatively-charged carriers at a point where the charge density is ρ and the density gradient $d\rho/dr$ is in the direction r . Consider two parallel planes distance δr apart, perpendicular to the direction r . The carriers in the vicinity of the two planes may be considered as a gas and to exert a partial pressure P which is known from statistical mechanics to be given by⁷

$$P = \frac{1}{3} n m \bar{v}^2 = -\frac{1}{3e} \rho m \bar{v}^2 \quad \dots \dots \dots (8)$$

where n is the density of carriers of mass m and \bar{v}^2 is the mean square thermal velocity. The net force F per unit area on the carriers between the planes is therefore in the r -direction and is

$$F = -\frac{1}{3e} \rho m \bar{v}^2 + \frac{1}{3e} \left(\rho + \frac{d\rho}{dr} \delta r \right) m \bar{v}^2 \quad \dots \dots (9)$$

$$= \frac{1}{3} m \bar{v}^2 \frac{1}{e} \frac{d\rho}{dr} \delta r \quad \dots \dots \dots (10)$$

This result may be written

$$F = \frac{2}{3} e \bar{W} \frac{1}{e} \frac{d\rho}{dr} \delta r \quad \dots \dots \dots (11)$$

where $e\bar{W} = \frac{1}{2} m \bar{v}^2$ is the mean thermal kinetic energy of the carrier. The number of carriers contained between the planes per unit area is $n\delta r$; hence the average force exerted on one carrier is $f = F/n\delta r = -eF/\rho\delta r$:

$$f = -\frac{2}{3} e \bar{W} \frac{1}{\rho} \frac{d\rho}{dr} \quad \dots \dots \dots (12)$$

This may be considered an average force exerted on a carrier due to a density gradient, to which, in general, must be added the force due to the potential gradient, so that the total force f_t on a carrier is

$$f_t = e \frac{dV}{dr} - \frac{2}{3} e \bar{W} \frac{1}{\rho} \frac{d\rho}{dr} \quad \dots \dots \dots (13)$$

$$= e \nabla V - \frac{2}{3} e \bar{W} \frac{1}{\rho} \nabla \rho \quad \dots \dots \dots (14)$$

in which the direction of the total force is implicit. As described above, the velocity does not increase indefinitely because of collisions. If ∇V and $\nabla \rho$ do not change during a free time t , a carrier will travel a distance $s = \frac{1}{2} a t^2$, where a is the acceleration due to f_t . The mean distance travelled by many carriers is $\bar{s} = \frac{1}{2} a \bar{t}^2$, which for an exponential distribution of free times can be shown⁸ to be $\bar{s} = a \bar{t}^2$. Hence the mean velocity averaged over many collisions is $v = \bar{s}/\bar{t} = a \bar{t}$. With use of the relation $f_t = ma$, the mean velocity is

$$v = \frac{\bar{t}}{m} f_t = \frac{\bar{t}}{m} \left(e \nabla V - \frac{2}{3} e \bar{W} \frac{1}{\rho} \nabla \rho \right) \quad \dots \dots (15)$$

which may be written

$$v = \mu \nabla V - D \frac{1}{\rho} \nabla \rho \quad \dots \dots \dots (16)$$

where the term in ∇V is the drift component and that in $\nabla \rho$ is the diffusion component of the velocity. By definition,

$$\mu \equiv e\bar{i}/m \quad \dots \quad (17)$$

$$D \equiv \frac{2}{3} \bar{W} \mu \quad \dots \quad (18)$$

Eqn. (16) is the fourth relation between the variables i , ρ , v and V , this time applicable when the device dimensions are large compared with the mean free path of the carriers (e.g. to electrons and positive ions in a plasma, to negative and positive ions in an electrolyte, to electrons in a metal, and to electrons and holes in a semiconductor). A more useful relationship may be obtained between μ and D than that given by eqn. (18) in special cases: electrons in a metal obey Fermi-Dirac statistics, for which the mean kinetic energy is given by⁹

$$\bar{W} = \frac{3}{5} W_f \quad \dots \quad (19)$$

where W_f is the height of Fermi level above the bottom of the conduction band. Hence, for metals,

$$\frac{D}{\mu} = \frac{2}{5} W_f \quad \dots \quad (20)$$

In all the other applicable cases listed above the carriers usually obey Maxwell-Boltzmann statistics, for which the mean energy is given by¹⁰

$$\bar{W} = \frac{3}{2} \frac{kT}{e} \quad \dots \quad (21)$$

Hence, for plasmas, electrolytes and semiconductors,

$$\frac{D}{\mu} = \frac{kT}{e} \quad \dots \quad (22)$$

This result is known as Einstein's relation, and was first applied to electrolytes.

The four equations necessary to solve for i , ρ , v and V are thus in general given by eqns. (1)-(3) and eqns. (7) or (16) as appropriate:

$$i = \rho v + \epsilon \frac{\partial E}{\partial t} \quad \dots \quad (23)$$

$$\nabla^2 V = -\rho_e / \epsilon \quad \dots \quad (24)$$

$$\frac{\partial \rho}{\partial t} = e(g - r) - \nabla \cdot (\rho v) \quad \dots \quad (25)$$

$$v = \begin{cases} \sqrt{\frac{2e}{m}} V^{1/2} & \text{free acceleration} \quad \dots \quad (26a) \\ \mu V & \text{drift} \quad \dots \quad (26b) \\ -D \frac{1}{\rho} \nabla \rho & \text{diffusion} \quad \dots \quad (26c) \end{cases}$$

together with the subsidiary relation

$$\rho_e = \rho_p + \rho_n + \rho_d + \rho_a \quad \dots \quad (27)$$

It is convenient to make immediately two basic approximations. First, the generation and recombination term in eqn. (25) is ignored; in semiconductors and plasmas this amounts to saying that in the time it takes carriers to cross the device there is negligible change in density from these causes; for electrons in vacuo no approximation is involved, since generation and recombination cannot occur, and so this term is identically zero. Second, the time-dependent terms in eqns. (23) and (25) are ignored. This implies the condition that in the time it takes

carriers to cross the device there is negligible change in the density or in the field; under these assumptions, eqns. (23) and (25) may be combined, and under the further restriction of one-dimensional current flow the basic equations reduce to

$$i = \rho v \text{ independent of position} \quad \dots \quad (28)$$

$$\frac{d^2 V}{dx^2} = -\frac{\rho_e}{\epsilon} \quad \dots \quad (29)$$

$$v = \begin{cases} \sqrt{\frac{2e}{m}} V^{1/2} & \text{free acceleration} \quad \dots \quad (30a) \\ \mu \frac{dV}{dx} & \text{drift} \quad \dots \quad (30b) \\ -D \frac{1}{\rho} \frac{d\rho}{dx} & \text{diffusion} \quad \dots \quad (30c) \end{cases}$$

together with the subsidiary relation

$$\rho_e = \rho_p + \rho_n + \rho_d + \rho_a \quad \dots \quad (31)$$

The basic equations of current flow will be applied to some specific structures after some general properties of charge-controlled devices are discussed in the next Section.

(3) CHARGE-CONTROL DEVICES

Johnson and Rose¹ have shown how the basic principles of many vacuum and semiconductor devices may be described very simply in terms of charge-control relations. Some of their results are here summarized and generalized.

The principle of a charge-controlled device is that of control of the current between two output electrodes by means of the charge placed on a third electrode, the gate. This simple and general arrangement is shown in Fig. 1, where the geometry of

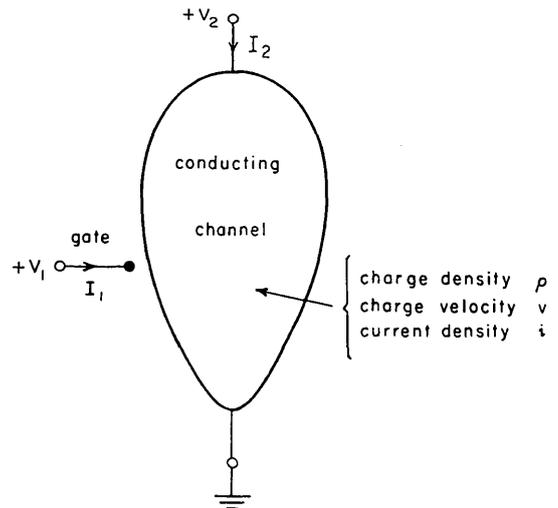


Fig. 1.—Arbitrary geometry and polarity conventions for a charge-controlled device.

The output current I_2 in the conducting channel is controlled by the charge on the gate.

the conducting path between the output electrodes is arbitrary, and in which the gate is shown schematically in order to avoid, for the present, a discussion of its physical structure. For convenience, one output electrode is earthed and may also be considered the reference terminal for the gate. The following relation holds between the output current density I_2 , the total charge in transit Q (per unit cross-section of base), and the

average transit time τ_t of the carriers between the output electrodes:

$$I_2 = Q/\tau_t \quad \dots \quad (32)$$

This is a relation of fundamental importance.

The prime purpose of the device is to provide a means of controlling the output current. This is usually most conveniently accomplished by controlling the charge in transit. The rate of change of I_2 with Q is then a quantity of interest:

$$\frac{dI_2}{dQ} = \frac{1}{\tau_t} \left(1 - \frac{Q}{\tau_t} \frac{d\tau_t}{dQ} \right) \quad \dots \quad (33)$$

In general, τ_t is not independent of Q , and, in fact, a definite functional relationship exists for each of the three types of current discussed in the previous Section. This relationship may be established as follows. The four quantities i , ρ , v and V can in principle be determined completely as functions of position in a device of arbitrary but fixed geometry and subject to a known set of applied voltages or currents. If the electrical boundary conditions are changed in such a way that the charge density is increased at all points by some factor, the total charge is increased by the same factor. Hence

$$Q \propto \rho \quad \dots \quad (34)$$

Similarly, if the electrical boundary conditions are changed so that the charge velocity is increased at all points by some factor, the average transit time is decreased by the same factor. Hence

$$\tau_t \propto 1/v \quad \dots \quad (35)$$

Likewise, relations can be found between v and V from eqn. (26):

$$v \propto \begin{cases} V^{1/2} & \text{free acceleration} \\ V & \text{drift} \\ \text{const.} & \text{diffusion} \end{cases} \quad \dots \quad (36)$$

From eqn. (24), $\rho_e \propto V$, and if the net charge density ρ_e is made up entirely of the charge density of the carrier whose current is being considered, then

$$\rho \propto V \quad \dots \quad (37)$$

Replacement of ρ_e by ρ is exact for electrons in vacuo because ρ_p , ρ_d and ρ_a are each zero, but is only approximate for space-charge-limited flow of carriers in a semiconductor. Combining the above relations gives

$$\tau_t \propto \begin{cases} \frac{1}{Q^{1/2}} & \text{space-charge-limited free acceleration current} \\ \frac{1}{Q} & \text{space-charge-limited drift current} \\ \text{const.} & \text{diffusion current} \end{cases} \quad (38)$$

Hence a general functional relationship exists between τ_t and Q such that

$$\tau_t \propto \frac{1}{Q^n} \quad \dots \quad (39)$$

where $n = 0$ for diffusion current, 0.5 for space-charge-limited free-acceleration current, and 1 for space-charge-limited drift current. It then follows from eqn. (33) that in a charge-controlled device

$$\frac{dI_2}{dQ} = \frac{1+n}{\tau_t} \quad \dots \quad (40)$$

The next step is to determine how the charge may be changed. This is the function of the gate, which is such that a change in

the gate charge, q , gives rise to an equal and opposite change in Q (in the absence of charge multiplication effects such as avalanche breakdown). In general, a change in gate (input) potential V_1 will be required to change the gate charge q .

Define

$$C_1 = \left. \frac{\partial q}{\partial V_1} \right|_{V_2} \quad \dots \quad (41)$$

as the incremental input capacitance, which may or may not be dependent on V_1 . Similarly, define

$$C_2 = \left. \frac{\partial Q}{\partial V_2} \right|_{V_1} \quad \dots \quad (42)$$

as the incremental output capacitance, which may or may not be dependent on V_2 . The dependence of the output current on the input and output voltages may now be found as

$$g_m = \left. \frac{\partial I_2}{\partial V_1} \right|_{V_2} = \frac{\partial I_2}{\partial Q} \left. \frac{\partial Q}{\partial V_1} \right|_{V_2} = \frac{(1+n)C_1}{\tau_t} \quad (43)$$

$$G_2 = \left. \frac{\partial I_2}{\partial V_2} \right|_{V_1} = \frac{\partial I_2}{\partial Q} \left. \frac{\partial Q}{\partial V_2} \right|_{V_1} = \frac{(1+n)C_2}{\tau_t} \quad (44)$$

where the one-to-one relationship between changes in q and Q is employed. The input admittance may contain a conductive as well as a capacitive component, defined as

$$G_1 = \left. \frac{\partial I_1}{\partial V_1} \right|_{V_2} \quad \dots \quad (45)$$

but explicit expressions for G_1 must await discussion of specific structures.

It must be emphasized that all the above five circuit parameters have been defined as differential coefficients, and thus represent incremental elements whose values depend on the particular operating point. An incremental circuit model may be constructed as shown in Fig. 2. It should also be borne in mind

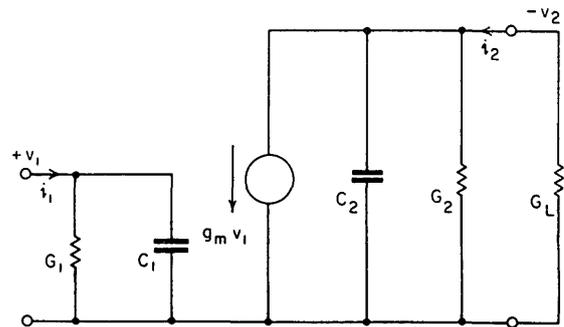


Fig. 2.—Basic incremental model of a charge-controlled device.

that the five parameters have been defined as short-circuit quantities, and hence part of the admittances G_1 , C_1 and G_2 , C_2 may be bridged directly between input and output rather than directly to earth. This possibility will be considered further in specific cases, but is neglected at present to avoid obscuring the fundamental properties. It may further be noted that the incremental model shown in Fig. 2 contains only intrinsic elements, i.e. those directly related to the charge in transit, and does not include extrinsic elements such as direct inter-electrode capacitance, ohmic lead resistance or leakage conductance.

The transconductance, g_m , represents the essential property of the device. From eqn. (43) it is seen that a large value of C_1 is necessary for a large transconductance, and the input capacitance is therefore an integral part of the device. On the other

hand, the input conductance, G_1 , merely represents a loss of input current, and C_2 and G_2 merely represent a loss of output current; hence these elements may be considered as 'parasitic' in contrast with the 'essential' elements g_n and C_1 .

Some incremental circuit properties of the charge-controlled device for sinusoidal signals are of interest. The current gain, G_i , in the presence of a load conductance G_L is

$$G_i = g_m \left(\frac{1}{G_1 + j\omega C_1} \right) \left(\frac{G_L}{G_2 + j\omega C_2 + G_L} \right) \quad (46)$$

A special case of interest is the low-frequency short-circuit current gain β , defined by

$$\beta = G_i |_{G_L \rightarrow \infty, \omega \ll G_1/C_1} = \frac{g_m}{G_1} \quad (47)$$

The voltage gain, G_v , in the presence of a load conductance G_L is

$$G_v = \frac{g_m}{G_2 + j\omega C_2 + G_L} \quad (48)$$

One special case of interest is the low-frequency open-circuit voltage gain μ , defined by

$$\mu = G_v |_{G_L \rightarrow 0, \omega \ll G_2/C_2} = \frac{g_m}{G_2} = \frac{C_1}{C_2} \quad (49)$$

by eqns. (43) and (44). Another special case of interest is the low-frequency voltage gain when the load conductance is large:

$$G_v |_{G_L \gg G_2, \omega \ll G_L/C_2} = \frac{g_m}{G_L} \quad (50)$$

It has been shown above how the small-signal circuit properties of a charge-controlled device can be expressed very simply in terms of a few fundamental relations. Explicit equations for the circuit parameters may be obtained by application of the basic current-flow equations of Section 2 to specific device types and geometries. The following Sections indicate how the important characteristics of vacuum tubes and transistors may be determined by this approach.

(4) VACUUM DEVICES

The fundamental characteristics of low-frequency vacuum tubes can be simply derived by consideration, first of the properties of a metal-vacuum junction (thermionic emission), and second of the application of the basic current-flow equations to electrons in vacuo. Finally, application of the basic charge-control relations leads to equations for the elements in the incremental model.

(4.1) The Vacuum Diode

Electrons in random thermal motion in a metal constitute a random thermal current. On the average, of course, the net current is zero. However, it is frequently convenient to consider the unidirectional random thermal current, i.e. the thermal current in one direction which in thermal equilibrium is balanced by an equal current in the opposite direction. Since such electrons possess a range of energies whose distribution is expressed by the Fermi-Dirac distribution function, it is also convenient to consider the unidirectional thermal current which is due only to electrons with energies above some specified value W_0 . If this value is appreciably above the Fermi level W_f , electrons above W_0 obey Maxwell-Boltzmann statistics to a good approximation, and it can then be shown that the unidirectional thermal current, I_s , due to electrons with energies above W_0 is given by¹¹

$$I_s = AT^2 e^{-W_0/W_i} \quad (51)$$

where A is a coefficient of value $1.20 \times 10^6 \text{ amp/m}^2/\text{deg K}^2$, $W_\phi = |W_0 - W_f|$, and $W_i = kT/e$.

The electron current emitted from the surface of a metal can now be found by setting W_ϕ in eqn. (51) equal to the work function, ϕ_m , of the material (Fig. 3), i.e. the thermionic emission

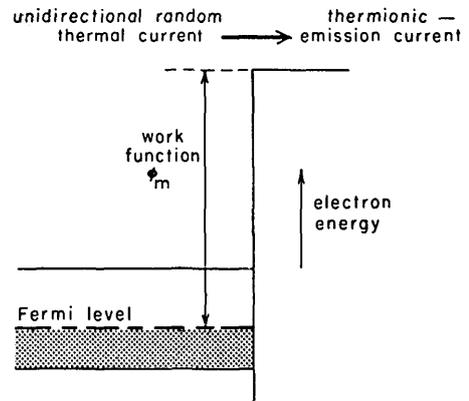


Fig. 3.—Energy diagram at the surface of a metal.

current is merely the unidirectional random thermal current of electrons in the metal with energies sufficient to overcome the work function. However, because of various effects due to the major crystal discontinuity at the surface, the numerical value of A in eqn. (51) varies rather widely from one material to another. Since practical values of the work function exceed 1 eV, it is necessary to heat the metal to obtain an emission current of a magnitude useful for electronic device purposes.

It must be remembered that the current given by eqn. (51) is the unidirectional current, and that the net current is the algebraic sum of this and any current which may flow in the opposite direction. Therefore the actual emission current from a metal surface is only equal to I_s if no electrons are returned to the surface: this requires an electric field external to the metal to draw away electrons as they are emitted. Moreover, the electric field must be sufficiently strong to penetrate the negative space-charge set up by the emitted electrons. If this condition is not fulfilled, electrons will be returned to the metal by repulsion from the space charge of those already emitted, and the net current from the surface will be less than I_s . An expression for the net current emitted may be found by application of the basic current equations (electron flow in vacuo). For this purpose an accelerating potential must be applied, and the simplest geometry to consider is that of a parallel-plate diode as shown

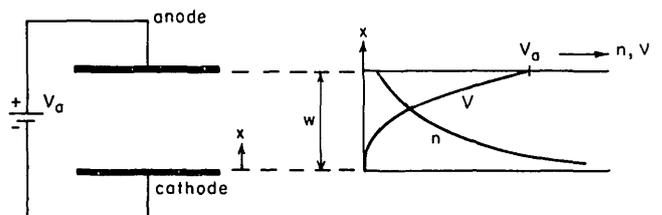


Fig. 4.—Geometry of a parallel-plate vacuum diode, and potential and electron density distributions in space-charge-limited current flow.

in Fig. 4, where the anode with accelerating voltage V_a is a distance w from a heated emitting cathode.

The equations of current flow applicable to electrons in a vacuum are included as special cases of eqns. (28)–(31). For

parallel-plate (one-dimensional) geometry, and for $\rho_p = \rho_d = \rho_a = 0$, the reduced equations are

$$i = -I = -env \quad \text{independent of } x \quad \dots (52)$$

$$-\frac{dE}{dx} = \frac{d^2V}{dx^2} = \frac{en}{\epsilon_0} \quad \dots (53)$$

$$v = \sqrt{\frac{2e}{m}} V^{1/2} \quad \dots (54)$$

where $\rho_e = \rho_n = -en$. It is convenient to write $I = -i$ since the conventional current is positive in the $-x$ direction in Fig. 4. The method of solution recently reintroduced by Moullin¹² is most appropriate from the charge-control aspect: the electron density, n , is eliminated between eqns. (52) and (53), and then with $v = dx/dt$, integration of the result leads to

$$E = -\frac{I}{\epsilon_0} t \quad \dots (55)$$

where the usual first-order approximation of zero electric field at the cathode is made. Now, instead of using eqn. (54) it is more convenient to employ directly the fundamental relation upon which eqn. (54) is based, namely Newton's law

$$-eE = m \frac{d^2x}{dt^2} \quad \dots (56)$$

Eqns. (55) and (56) then give

$$\frac{d^2x}{dt^2} = \frac{e}{m} \frac{I}{\epsilon_0} t \quad \dots (57)$$

which, as Moullin has pointed out, implies that the accelerating force on an electron in transit may be considered as being due only to the total electron charge behind that electron.

Direct integration of eqn. (57), with the other usual first-order approximation of zero initial velocity at the cathode, leads to the following expressions for the velocity and position of an electron as functions of time:

$$\frac{dx}{dt} = \frac{e}{2m} \frac{I}{\epsilon_0} t^2 \quad \dots (58)$$

$$x = \frac{e}{6m} \frac{I}{\epsilon_0} t^3 \quad \dots (59)$$

The transit time of electrons from cathode to anode is then given by putting $x = w$ in eqn. (58), as

$$\tau_t = \left(\frac{6m\epsilon_0}{e}\right)^{1/3} \frac{w^{1/3}}{I^{1/3}} \quad \dots (60)$$

By making use of the basic charge-control relation $Q = I\tau_t$, the transit time and the current can each be expressed in terms of the total charge in transit Q , i.e.

$$\tau_t = \left(\frac{6m\epsilon_0}{e}\right)^{1/2} \frac{w^{1/2}}{Q^{1/2}} \quad \dots (61)$$

$$I = \left(\frac{e}{6m\epsilon_0}\right)^{1/2} \frac{Q^{3/2}}{w^{1/2}} \quad \dots (62)$$

It may be noted that eqn. (61) corroborates the general functional relationship between τ_t and Q developed in Section 3. The anode potential V_a can also be found in terms of Q through eqn. (54), where the velocity at the anode is obtained by putting $t = \tau_t$ in eqn. (58):

$$V_a = \frac{3}{4\epsilon_0} wQ \quad \dots (63)$$

Finally, elimination of Q between eqns. (62) and (63) leads to the familiar three-halves power-law relation between I and V_a :

$$I = \frac{4\epsilon_0}{9} \sqrt{\frac{2e}{m}} \frac{V_a^{3/2}}{w^2} \quad \dots (64)$$

The 2-electrode vacuum diode is a degenerate form of charge-control device in which the input and output terminals are the same, namely the anode. Hence in calculating incremental parameters from the charge-control formulae of Section 3, subscripts 1 and 2 can be omitted and thus

$$C = \frac{\partial Q}{\partial V_a} = \frac{4\epsilon_0}{3w} \quad \dots (65)$$

from eqns. (42) and (63), and

$$G = \frac{\partial I}{\partial V_a} = \frac{(1+n)C}{\tau_t} \quad \dots (66)$$

$$= \frac{2\epsilon_0}{w\tau_t} \quad \dots (67)$$

from eqn. (44), since $n = 0.5$ for free-acceleration electron motion in vacuo. The incremental conductance G may be expressed in terms of either the anode current or the anode voltage by eqn. (61) and either eqn. (62) or (63):

$$G = \left(\frac{4e\epsilon_0^2}{3m}\right)^{1/3} \frac{I^{1/3}}{w^{4/3}} = \frac{2\epsilon_0}{3} \left(\frac{2e}{m}\right)^{1/2} \frac{V_a^{1/2}}{w^2} = \frac{3}{2} \frac{I}{V_a} \quad (68)$$

It may be noted that the incremental conductance increases as the cube root of the operating current, while the incremental capacitance is constant at four-thirds times the electrostatic value in the absence of space charge.

(4.2) The Vacuum Triode

In order to convert the 2-electrode vacuum diode into a 3-electrode control device, some means must be found of controlling the electron charge in transit between cathode and anode, i.e. a gate structure should be introduced in the electron path. Ideally, the charge placed on the gate should be the sole controlling agent and the gate should draw no current itself. Under these conditions the input conductance would be zero, and the anode voltage would have no effect on the charge in transit, thus making the output conductance and capacitance both zero.

Since in a parallel-plate diode the cathode current is determined by the potential and position of a unipotential plane (the anode), the ideal gate structure would be a unipotential plane. However, in order that the gate should draw no current, the gate should be permeable to electrons so that all the cathode current could pass through the gate and reach the anode. A metallic layer a few atoms thick (less than the mean free path of electrons in the metal) would fulfil these requirements. Although a physical structure of this type is not practical, it is of interest to derive the properties of such an ideal triode.

Let an ideal gate be placed a distance d_1 from a planar cathode, and let an anode be placed a distance d_2 from the gate, as shown in Fig. 5. If the gate is given a positive potential V_a , the expressions relating current, voltage, charge density and transit time between cathode and gate are given by the same equations as for the vacuum diode—eqns. (161)–(164)—with w replaced by d_1 . By assumption, all this current passes through the gate. If the accelerating electric field in the gate-anode region is sufficiently high, the current between the gate and the anode will be emission-limited to the amount determined by the cathode-gate condi-

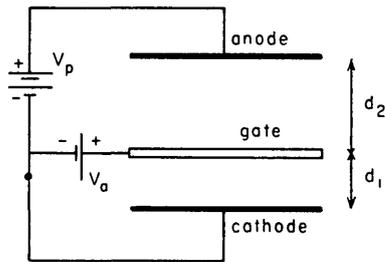


Fig. 5.—Geometry of a parallel-plate ideal vacuum triode. The gate is an equipotential plane at potential V_a , but is permeable to electrons.

tions. Thus, space charge in the gate-anode region may be neglected, and also the transit time will be small compared with that between the cathode and the gate. Thus, so long as the anode potential is sufficiently high, the anode current will be independent of anode voltage and the total charge in transit and the transit time between cathode and anode are essentially the same as those between cathode and gate given by eqns. (61)–(64), again with w replaced by d_1 .

The incremental properties of this ideal triode may now be found directly by application of eqns. (40)–(45), where gate and anode quantities are distinguished by subscripts 1 and 2, respectively:

$$C_1 \equiv \left. \frac{\partial Q}{\partial V_1} \right|_{V_2} = \frac{4\epsilon_0}{3w} \dots \dots \dots (69)$$

$$C_2 \equiv \left. \frac{\partial Q}{\partial V_2} \right|_{V_1} = 0 \dots \dots \dots (70)$$

$$g_m = \frac{(1+n)C_1}{\tau_t} = \left(\frac{4e\epsilon_0^2}{3m} \right)^{1/3} \frac{I^{1/3}}{d_1^{4/3}} = \frac{2\epsilon_0}{3} \left(\frac{2e}{m} \right)^{1/2} \frac{V_1^{1/2}}{d_1^2} = \frac{3}{2} \frac{I}{V_1} \dots \dots \dots (71)$$

$$G_2 \equiv \frac{(1+n)C_2}{\tau_t} = 0 \dots \dots \dots (72)$$

$$G_1 \equiv \left. \frac{\partial I_1}{\partial V_1} \right|_{V_2} = 0 \dots \dots \dots (73)$$

Hence the incremental model of an ideal triode contains only the 'essential' elements C_1 and g_m discussed in Section 3 (apart from direct electrostatic capacitances).

Since the ideal gate structure of a metallic layer a few atoms thick is impractical, the nearest realizable approach is to gather the atoms together into large groups with spaces between them. This corresponds to the usual grid structure. The performance will, of course, no longer be ideal, since the potential in the plane of the grid wires will contain ripples, so that the potential is that of the grid at a wire but is influenced strongly by the anode potential in between the grid wires. For purposes of analysis it is convenient to smooth out these ripples and to suppose that the grid plane is at an average potential whose value is between that of the grid and that of the anode. This approximation will be reasonably accurate so long as the potential ripples are negligible at the cathode—a condition adequately met if the grid-cathode distance is several times the spacing between the grid wires. Thus the cathode current may be considered to be determined by the potential of the 'equivalent grid plane', whose potential may be determined by the electrostatic relationships depicted in Fig. 6. The equivalent-grid-plane concept has been described by Thompson¹³ and extensively

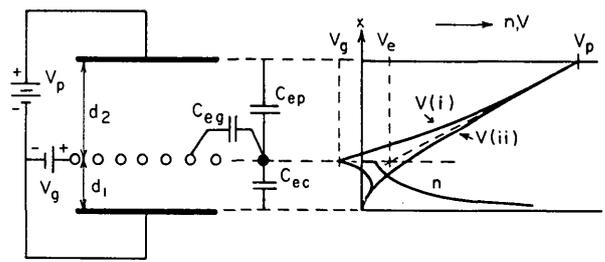


Fig. 6.—Geometry of a vacuum triode, and potential and electron density distributions. Potential $V(i)$ is that on a line through a grid wire; $V(ii)$ is that on a line between grid wires.

discussed by Dow.¹⁴ The potential, V_e , of the equivalent grid plane may be found, in terms of the three capacitances shown in Fig. 6, to be

$$V_e = \frac{C_{eg} V_g + V_p}{1 + \frac{C_{eg} + C_{ec}}{C_{ep}}} \dots \dots \dots (74)$$

Now the capacitances C_{ep} and C_{ec} are just the parallel-plate capacitances appearing between the two pairs of planes, and are given by

$$C_{ep} = \frac{\epsilon_0}{d_2} \dots \dots \dots (75)$$

$$C_{ec} = \frac{4}{3} \frac{\epsilon_0}{d_1} \dots \dots \dots (76)$$

where due note is taken of the presence of space charge between the cathode and grid and its absence between the grid and anode. The third capacitance, C_{eg} , is much more difficult to calculate, since it depends on the geometry of the actual grid structure. However, it may be noted that this capacitance always occurs in the ratio C_{eg}/C_{ep} , and the labour of calculating C_{eg} may be avoided by defining this ratio as a geometry-dependent parameter μ , thus

$$\mu \equiv C_{eg}/C_{ep} \dots \dots \dots (77)$$

With help of eqns. (75)–(77), eqn. (74) becomes

$$V_e = \frac{\mu V_g + V_p}{1 + \mu + \frac{\delta}{3} \frac{d_2}{d_1}} \equiv \delta V_g + \frac{\delta}{\mu} V_p \dots \dots (78)$$

where δ is again a geometry-dependent parameter defined in the above equation. It may be noted that if $C_{eg} \rightarrow \infty$, then $\mu \rightarrow \infty$ and $\delta \rightarrow 1$, thus $V_e \rightarrow V_g$ and the grid becomes ideal.

The current flow is determined by the same relations as in the vacuum diode given by eqns. (61)–(64), where V_a is replaced by V_e and w is replaced by d_1 . If cathode current is to flow, V_e must be positive, but if grid current is not to flow, V_g must be negative. This is the normal condition of operation. In the absence of grid current the anode current is

$$I = \frac{4\epsilon_0}{9} \left(\frac{2e}{m} \right)^{1/2} \frac{V_e^{3/2}}{d_1^2} = \frac{4\epsilon_0}{9} \left(\frac{2e}{m} \right)^{1/2} \frac{\delta^{3/2}}{d_1^2} \left(V_g + \frac{1}{\mu} V_p \right)^{3/2} \dots (79)$$

which is the familiar three-halves power law for a triode.

The incremental parameters may be found by applying the basic charge-control relationships of eqns. (40)–(45), where again subscripts 1 and 2 refer to grid anode and:

$$C_1 \equiv \left. \frac{\partial Q}{\partial V_1} \right|_{V_2} = \frac{\partial Q}{\partial V_e} \frac{\partial V_e}{\partial V_1} \Big|_{V_2} = \frac{4}{3} \frac{\epsilon_0}{d_1} \delta \dots \dots (80)$$

$$C_2 \equiv \left. \frac{\partial Q}{\partial V_2} \right|_{V_1} = \left. \frac{\partial Q}{\partial V_e} \frac{\partial V_e}{\partial V_2} \right|_{V_1} = \frac{4}{3} \frac{\epsilon_0}{d_1} \frac{\delta}{\mu} \dots (81)$$

$$g_m = \frac{(1+n)C_1}{\tau_i} = \delta \left(\frac{4e\epsilon_0^2}{3m} \right)^{1/3} \frac{I^{1/3}}{d_1^{4/3}} \dots (82)$$

$$G_2 = \frac{(1+n)C_2}{\tau_i} = \frac{\delta}{\mu} \left(\frac{4e\epsilon_0^2}{3m} \right)^{1/3} \frac{I^{1/3}}{d_1^{4/3}} = \frac{g_m}{\mu} \dots (83)$$

$$G_1 = 0 \text{ for negative grid} \dots (84)$$

It should be remembered that the parameters derived above are the intrinsic parameters only, i.e. those for which the charge in transit is directly responsible. In addition to these elements, any extrinsic elements which may be present should be included in the incremental model, and the most important of these is the direct anode-grid inter-electrode capacitance C_{12} . This is easily found by star-delta transformation of the three capacitances shown in Fig. 6 to be

$$C_{12} = \frac{\epsilon_0}{d_2} \delta \dots (85)$$

By the same transformation, it is easily shown that

$$\mu \equiv C_{eg}/C_{ep} = C_1/C_2 \dots (86)$$

and is independent of current so long as the restrictions imposed on the above derivation are valid—a well-known result which may be contrasted with the operating-current dependence of both g_m and G_2 as shown by eqns. (82) and (83).

For typical values of μ between 10 and 100, the parameter δ is essentially unity and it may be observed that the transconductance, g_m , of the practical triode is almost as great as that of the ideal triode. From a circuit point of view the principal effect of the departure from ideality in a practical triode is to introduce a finite (and quite large) output conductance G_2 .

(4.3) Vacuum Tetrodes and Pentodes

Although historically the screen grid was introduced¹⁵ in order to decrease the extrinsic element C_{12} , in retrospect the extra grid may be considered as serving the more fundamental purpose of making the practical triode closer to the ideal triode. This is accomplished by making the cathode current less dependent on the anode voltage, which implies a larger value of μ and hence a smaller value of G_2 . Just as the grid and anode were represented by an equivalent potential plane which allowed the triode to be reduced to an equivalent diode, so the screen grid and anode can be represented by an equivalent potential plane which allows the tetrode to be reduced to an equivalent triode. The geometry is such that the potential of the second equivalent plane is determined largely by the screen potential, the anode potential having little effect. Thus the anode potential of the equivalent diode is, as it were, twice removed from the actual anode potential.

One additional process of reduction allows the pentode to be represented by an equivalent diode, and the routine may be extended to a greater number of grids. Reference may be made to Thompson¹³ and Dow¹⁴ for further details. From a charge control point of view, no new principles are introduced by extension to multi-electrode tubes, and the subject will not be further pursued here.

(5) METAL-SEMICONDUCTOR DEVICES

The fundamental characteristics of metal-semiconductor devices can be derived by consideration, first of the properties of metal-semiconductor junctions, and second, of the application of the basic current-flow equations to carriers in semiconductors.

Only 2-terminal devices will be mentioned here and only highly idealized models will be employed, since the features of interest for the present discussion are not the characteristics of surface-barrier or point-contact transistors but rather the similarity of the carrier injection mechanism under certain conditions to that in vacuum tubes, and the properties of the space-charge region near the contact which are of importance in junction devices.

Consider a metal of work function ϕ_m and an n -type semiconductor of work function ϕ_s , where $\phi_m > \phi_s$. When the metal and the semiconductor are separate, the energy diagrams are as shown in Figs. 8(a) and (c). If the two materials are brought into contact, electrons will flow from the semiconductor to the metal, thus exposing the fixed positive lattice charges in the semiconductor and adding negative charge to the metal. The transfer of electrons causes the metal to become more negative and the semiconductor more positive, thus shifting the energy diagrams relative to one another, and the flow of electrons will cease when the two Fermi levels have become equal.

Consider now conditions in the semiconductor. The equation for net charge density ρ_e , eqn. (31), may be reduced to

$$\rho_e = \rho_n + \rho_d \dots (87)$$

where ρ_d is the net fixed charge density (donors minus acceptors) and ρ_n is the mobile electron charge density. If the material is strongly n -type, mobile positive charge due to holes may be neglected. If the semiconductor is isolated, the net charge density ρ_e is zero and $\rho_n = -\rho_d$. However, after equilibrium with the metal has been reached, ρ_n has become smaller in magnitude near the junction, so that net positive charge remains. Poisson's equation, (29), then requires a gradient of electric field and hence a curvature of potential within the semiconductor. The total change in potential within the semiconductor will depend on the magnitude of the net charge density and the distance over which it exists. A general solution is not obtainable in closed form, but a good approximation may be obtained by assuming that complete mobile electron depletion occurs throughout a distance w_s from the junction into the semiconductor, and that charge neutrality exists at all larger distances. This assumption may be justified by noting that when the junction is in thermal equilibrium the mobile electrons in the semiconductor will occupy a Maxwell-Boltzmann distribution in which the density falls off exponentially with increasing energy. Thus the potential in the semiconductor need depart only slightly from its original value for the electron density to change considerably, so that the net charge density is essentially that due only to the fixed charges. The assumption of a sudden discontinuity between regions of complete electron depletion and of charge neutrality also has the very considerable advantage that conditions in each region can be evaluated separately and then matched at the discontinuity.

Conditions in the metal-semiconductor junction near the discontinuity in the semiconductor are shown in Fig. 7. It is convenient to establish a co-ordinate system such that distance x is measured from the discontinuity towards the junction, and the charge-neutral region is considered as at zero potential. For one-dimensional geometry, Poisson's equation applied to the depletion region is

$$\frac{d^2V}{dx^2} = -\frac{\rho_e}{\epsilon} = -\frac{\rho_d}{\epsilon}, 0 < x < w_s \dots (88)$$

Two integrations of eqn. (88) then give the relation between the total change in potential V_s in the semiconductor and the distance w_s over which the change occurs, i.e.

$$V_s = \frac{\rho_d}{2\epsilon} w_s^2 \dots (89)$$

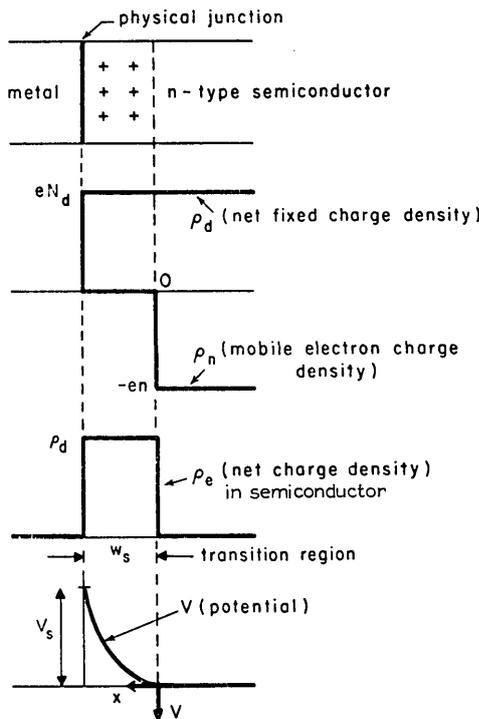


Fig. 7.—Charge densities, net charge density and potential in an *n*-type semiconductor near a metal contact.

A sharp discontinuity is assumed between a transition region of complete charge depletion and a region of charge neutrality.

A precisely analogous result is obtained for the metal, which becomes more negative by an amount V_m because this net charge density is negative owing to addition of electrons from the semiconductor. However, it is easily seen that V_m is much less than V_s , for the following reason. The total number of additional electrons in the metal cannot be greater than the total number which left the semiconductor, and because the original density of electrons in the metal was much greater than that in the semiconductor, the transfer of electrons from one to the other results in a much smaller fractional disturbance in their density in the metal than in the semiconductor. Consequently, a much smaller change in potential will occur in the metal than in the semiconductor. Since the junction between the metal and the semiconductor is in equilibrium when the two Fermi levels become equal, the total potential change across the junction is equal to the original difference between the work functions. By the above argument, the energy diagram of the metal will remain undisturbed up to the junction, and essentially the whole potential change will occur within the semiconductor. The results of the above discussion are shown in Fig. 8, where the energy diagrams of the metal and the semiconductor are isolated in Figs. 8(a) and (c), and the combined energy diagram after equilibrium has been established is Fig. 8(b). The total distance over which the potential changes is called the 'transition region', which is essentially all within the semiconductor, and is equal to w_s . From eqn. (89), the thickness, w_s , of the transition region is related to the original difference in the work functions and the semiconductor donor charge density, ρ_d , by

$$\frac{\rho_d}{2\epsilon} w_s^2 = V_s = \phi_m - \phi_s \quad (90)$$

If an external battery is connected across the junction, the current is determined by those electrons which have enough energy to overcome the potential barrier at the junction. The

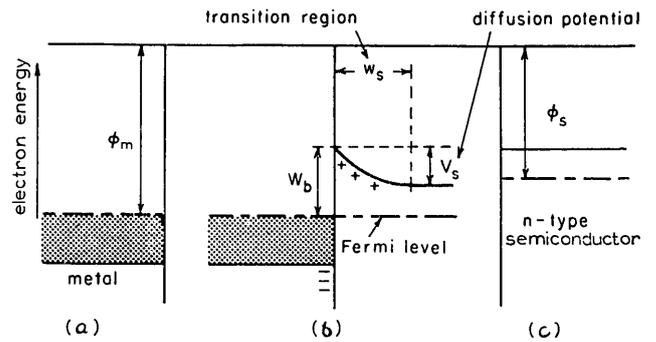


Fig. 8.—Energy diagrams of a metal and an *n*-type semiconductor, (a) and (c) before contact, and (b) in equilibrium after contact (rectifying contact, $\phi_m > \phi_s$).

situation is precisely the same as in a metal-vacuum junction, described in Section 4: the unidirectional thermal current across the junction is given by eqn. (51), in which W_ϕ is replaced by W_b (the barrier height) as defined in Fig. 8. Then

$$I_s = AT^2 e^{-W_b/W_t} \quad (91)$$

and A is again $1.20 \times 10^6 \text{ amp/m}^2/\text{deg K}^2$.

In thermal equilibrium, of course, the net junction current is zero, because there is on the average an equal and opposite current. However, if an external battery is connected in such a way that the semiconductor is made negative by a voltage V_b , the height of the barrier for electrons moving from the semiconductor will be reduced by the battery voltage while that for electrons moving from the metal to the semiconductor will be unaltered [see Fig. 9(a)]. Thus there will be a net flow of

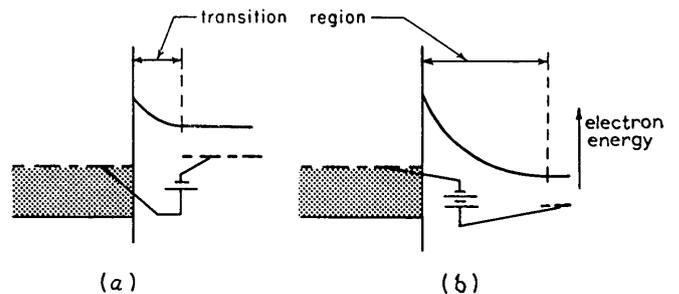


Fig. 9.—Dependence of transition-region thickness on applied voltage across a metal-to-*n*-type-semiconductor junction.

(a) Forward bias. (b) Reverse bias.

electrons from the semiconductor to the metal which constitutes a current equal to the difference of the two unidirectional thermal currents:

$$I = AT^2 e^{-(W_b - V_b)/W_t} - AT^2 e^{-W_b/W_t} = I_s (e^{V_b/W_t} - 1) \quad (92)$$

where the conventional current is positive when flowing from the metal to the semiconductor. If the semiconductor is made more positive, as shown in Fig. 9(b), the only difference in the above argument is that V_b reverses sign, and the barrier height is increased for electrons moving from the semiconductor to the metal. Thus in the limit the net current becomes merely the unidirectional thermal current I_s , and eqn. (92) is valid for either polarity of applied voltage. This is a rectifier characteristic. It should be noted that the above argument based on the unidirectional thermal current, known as the emission theory, is valid only if the thickness of the transition region, w_s , is smaller than the mean free path of electrons. If this condition does not hold, the current crossing the transition

region is controlled by drift and diffusion requirements, and the current/voltage relationship is of the same form as eqn. (92), except that I_s is different from the value given by eqn. (91). The argument leading to this result is known as the diffusion theory, and has been described by Spenke¹⁶ and by van der Ziel.¹⁷ The development will not be carried further, because it will instead be given in connection with semiconductor junctions.

Two further important results may be obtained from the preceding discussion. The total potential change V_s (known as the diffusion potential) across the structure determines the thickness of the transition region in a given material. When an external battery is connected, the total potential change is the algebraic sum of the diffusion potential (equal to the difference of the work functions) and the applied voltage. Under the condition that all the potential change is absorbed in the semiconductor, the thickness of the transition region, x_t , in the presence of a forward bias voltage V_b is given by eqn. (89) as

$$V_s - V_b = \frac{\rho_d x_t^2}{2\epsilon} \dots (93)$$

where $x_t = w_s$ when $V_b = 0$. The fact that the thickness of the transition region is dependent on the applied voltage is of great importance. For the present, the most important consequence of this property is that it gives rise to a transition-region capacitance, since a change in voltage changes the total net charge on one side of the junction. Since this capacitance is non-linear, it is usual to obtain an expression for the incremental transition-region capacitance C_t , thus: a change in voltage dV causes a change in transition region thickness dx_t , which implies a change in net charge of $\rho_n dx_t$. Hence

$$C_t \equiv \rho_n \frac{dx_t}{dV} = \frac{\epsilon}{x_t} = \sqrt{\frac{\rho_d \epsilon}{2}} \frac{1}{\sqrt{(V_s - V_b)}} \dots (94)$$

by eqn. (93).

The above description of the properties of a metal-to-*n*-type-semiconductor junction applies only when the work function of the metal is greater than that of the semiconductor. If the reverse is true, the potential within the semiconductor shown in Fig. 8 curves the other way, and the current resulting from an external applied potential is limited only by the resistance of the semiconductor. Such a structure is called an 'ohmic' or 'accumulation' junction. Analogous arguments can be applied to metal-to-*p*-type semiconductor junctions; such a junction is rectifying if the work function of the metal is smaller than that of the semiconductor, and ohmic if the reverse is true.

The brief discussion of metal-to-semiconductor junctions given in this Section has neglected the possible presence of surface double layers and surface states, which introduce greater complexity into the analysis and modify the results, in some cases to a drastic extent.^{16, 17}

(6) SEMICONDUCTOR DEVICES

The fundamental characteristics of semiconductor devices can be simply derived by application of the basic current-flow equations, first to the transition regions of *p-n* junctions (depletion regions), and then to the charge-neutral regions. Application of the basic charge-control relations leads to equations for the elements in the incremental model.

(6.1) The *p-n* Junction Diode

The energy diagrams of separate *n*-type and *p*-type semiconductors are shown in Fig. 10. When a junction between the two materials is formed, the energy diagrams shift relative to one another by an amount equal to the difference of the work functions. By analogy with the metal-to-semiconductor junction, essentially all the potential change occurs within the

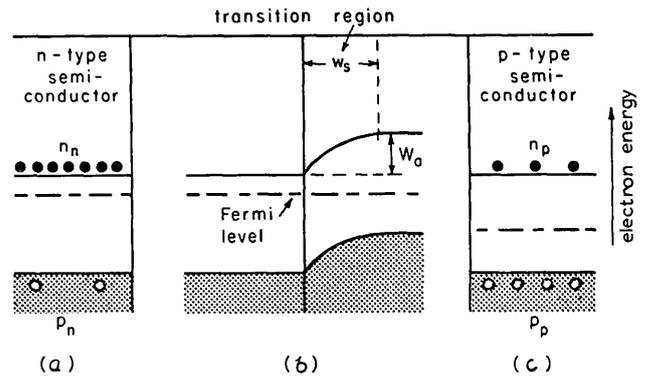


Fig. 10.—Energy diagram of *n*-type and *p*-type semiconductors, (a) and (c) before contact, and (b) in equilibrium after contact (conductivity of *n*-region much higher than that of *p*-region).

p-region if the conductivity of the *n*-region is much higher than that of the *p*-region. The energy diagram of the *n*-region is essentially undisturbed up to the junction, and a transition region of mobile charge depletion occurs in the *p*-region. The energy diagram of the *p-n* junction in equilibrium is shown in the middle of Fig. 10.

At first sight it appears that a *p-n* junction should be an ohmic structure, because both electrons and holes can easily cross the transition region. The ohmic resistance to electrons in the *p*-region is very high, because of the low density of electrons, and similarly the ohmic resistance to holes in the *n*-region is very high. Thus it seems that the device would exhibit a high resistance for either polarity of applied voltage. However, this argument does not account for the possibility of exchange of current between electrons and holes through the process of generation and recombination, nor for the possibility that minority carrier current may be carried by diffusion instead of by drift. Attempts to treat the *p-n* junction by the emission theory applicable to a metal-semiconductor junction also fail, because the thickness of the transition region is greater than the mean free path of mobile carriers. It is therefore necessary to apply the basic equations of current flow to the transition region and to the charge-neutral regions separately, and to obtain the complete solution by matching boundary conditions.

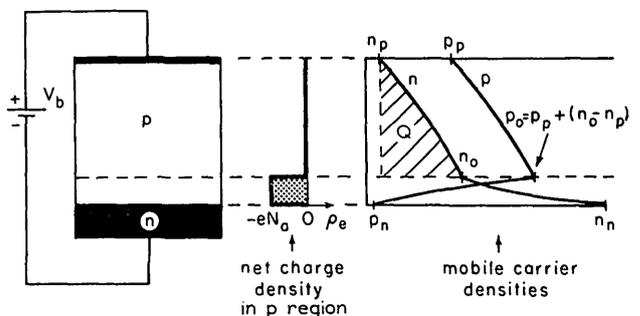


Fig. 11.—Geometry of a *p-n* junction diode, and electron and hole densities in the charge-depletion and the charge-neutral regions (high-level injection).

Consider a *p-n* junction diode as shown in Fig. 11. This is drawn in a manner similar to that of the vacuum diode of Fig. 4, where the *n*-region replaces the emitting cathode and the *p*-region replaces the vacuum. The *n*-region is assumed to be of very high conductivity, so that only conditions in the *p*-region need be considered. The simplifying assumption is made of complete mobile charge depletion in the transition region, thus exposing

the fixed acceptor charge density $\rho_a = -eN_a$, and of charge neutrality in the remainder of the p -region.

If an external battery of voltage V_b is applied across the junction with the p -type side positive, electrons will flow from the p -region through the n -region, thus increasing the electron density above its equilibrium value. Let the electron density at the edge of the transition region on the p -type side (the 'injected density') be n_0 in the presence of an externally-applied voltage V_b . The total voltage drop across the device, V_b , will in general be divided into two parts—that across the transition region, V_j , and that across the neutral p -region, V_p . (That across the n -region is neglected because of the high conductivity.) The voltage V_j subtracts from the diffusion potential W_a in the same way as shown for the metal-semiconductor junction in Fig. 9, and also affects the thickness of the transition region and gives rise to a capacitance.

A relation between the injected electron density and the voltage across the transition region may be obtained by applying the basic current-flow equations to the transition region. The electrons in the transition region are subject to opposing drift and diffusion tendencies. The net electron current which flows through the p - n junction device is the difference between these tendencies, and any reasonable net current represents only a very small unbalance of the opposing drift and diffusion tendencies. Thus it is a good approximation to suppose the drift and diffusion currents within the transition region to be equal and opposite even in the presence of an externally applied voltage.¹⁸ This condition is expressed for a one-dimensional structure by setting the sum of eqns. (30b) and (30c) equal to zero, where for electrons $\rho = -en$:

$$\mu_n \frac{dV}{dx} - D_n \frac{1}{n} \frac{dn}{dx} = 0 \quad \dots \quad (95)$$

Use of the Einstein relation, eqn. (22), then leads to

$$\frac{1}{W_t} dV - \frac{1}{n} dn = 0 \quad \dots \quad (96)$$

and then direct integration over the transition region gives

$$n_0 = n_n \epsilon^{-(W_a - V_j)/W_t} \quad \dots \quad (97)$$

since the electron density changes from n_n at the n -side to n_0 at the p -side with a corresponding change in potential of $W_a - V_j$. This result can be simplified by noting that in equilibrium when $V_j = 0$, n_0 is equal to its equilibrium value n_p , and hence W_a may be eliminated from eqn. (97), giving

$$n_0 = n_p \epsilon^{V_j/W_t} \quad \dots \quad (98)$$

This result is the familiar expression for the injected electron density into the charge-neutral part of the p -region.

A similar process leads to a relation between n_0 and the portion V_p of the external voltage which is dropped across the charge-neutral part of the p -region. If recombination in the p -region is neglected, the net hole current is zero at all points in the p -region, since none can flow into the n -region (unity emitter efficiency, because of the assumed very high n -region conductivity). Hence the sum of the hole drift and diffusion currents is zero, and

$$\mu_p \frac{dV}{dx} + D_p \frac{1}{p} \frac{dp}{dx} = 0 \quad \dots \quad (99)$$

The hole density at each end of the charge-neutral p -region is found from the subsidiary relation, eqn. (31), with $\rho_e = 0$. At the outside edge of the p -region the electron density is at its equilibrium value n_p , because of high surface recombination; hence the hole density is at its equilibrium value p_p . Thus

$$0 = p_p - n_p - N_a \quad \dots \quad (100)$$

At the edge of the transition region the electron density is n_0 , and if the hole density is p_0 , then

$$0 = p_0 - n_0 - N_a \quad \dots \quad (101)$$

Thus

$$p_0 = p_p + n_0 - n_p \simeq p_p + n_0 \quad \dots \quad (102)$$

if the p -region is fairly heavily doped. It is therefore known that the hole density varies from $p_p + n_0$ at one end of the charge-neutral p -region to p_p at the other, with a corresponding voltage change of V_p . Integration of eqn. (99) with these boundary values gives

$$p_p + n_0 = p_p \epsilon^{V_p/W_t} \quad \dots \quad (103)$$

The total applied voltage V_b may now be found from eqns. (98) and (103):

$$V_b = V_j + V_p = W_t \log_\epsilon \left[\frac{p_p Z(1 + Z)}{n_p} \right] \quad \dots \quad (104)$$

where

$$Z \equiv n_0/p_p \quad \dots \quad (105)$$

It remains to find the current through the p - n junction as a function of n_0 . Since the hole current is zero, the net current is essentially the electron current, which is the same at any cross-section, since recombination is neglected. It is most convenient to calculate the electron current in the charge-neutral part of the p -region, where the expressions for the electron and hole current densities are, from eqns. (28), (30b) and (30c),

$$i_n = -e\mu_n n \frac{dV}{dx} + eD_n \frac{dn}{dx} \quad \dots \quad (106)$$

$$i_p = 0 = -e\mu_p p \frac{dV}{dx} - eD_p \frac{dp}{dx} \quad \dots \quad (107)$$

In the charge-neutral p -region

$$p = p_p + n$$

at any point, and thus dV/dx and p can be eliminated from eqns. (106) and (107):

$$i_n = eD_n \frac{dn}{dx} \left(\frac{p_p - 2n}{p_p - n} \right) \quad \dots \quad (108)$$

Since i_n is independent of position, eqn. (108) can be integrated directly to give

$$\frac{i_n}{eD_n p_p} (x - w) = 2 \left(\frac{n - n_p}{p_p} \right) - \log_\epsilon \left(1 + \frac{n - n_p}{p_p} \right) \quad (109)$$

where the x -origin is chosen as the edge of the transition region on the p -type side, and the thickness of the charge-neutral p -region is w . Since $n = n_0$ when $x = 0$, eqn. (109) can be written

$$\frac{w}{eD_n p_p} I = 2 \left(Z - \frac{n_p}{p_p} \right) - \log_\epsilon \left(1 + Z - \frac{n_p}{p_p} \right) \quad \dots \quad (110)$$

where $I \equiv -i_n$ and is the total conventional current density flowing from the p -region.

It is of interest to notice from eqn. (109) that the electron density, n , is almost linear with distance in the charge-neutral p -region; hence the total electron charge in excess of the equilibrium value is closely given by

$$Q \simeq \frac{e(n_0 - n_p)w}{2} = \frac{ep_p w}{2} \left(Z - \frac{n_p}{p_p} \right) \quad \dots \quad (111)$$

and thus with use of the basic relation $I = Q/\tau_t$, the transit time, τ_t , of electrons is

$$\tau_t = \frac{w^2/2D_n}{2 - \frac{\log_\epsilon \left(1 + Z - \frac{n_p}{p_p} \right)}{Z - \frac{n_p}{p_p}}} \quad \dots \quad (112)$$

The carrier densities in the charge-neutral part of the p -region are shown in Fig. 11.

Eqns. (104) and (110)–(112) give the total voltage, total current, total charge and carrier transit time in terms of the physical constants of the device and of the parameter $Z = n_0/p_p$. This parameter is the same as the Z defined by Webster,¹⁹ and its value is a convenient means of distinguishing between the two limiting cases of low-level and high-level injection. If $Z \ll 1$, the injected electron density is much less than the equilibrium hole density in the p -region, and the parametric equations reduce to

$$V_b = W_t \log_e (n_0/n_p) \quad (113)$$

$$I = \frac{eD_n(n_0 - n_p)}{w} \quad (114)$$

$$Q = \frac{e(n_0 - n_p)w}{2} \quad (115)$$

$$\tau_t = w^2/2D_n \quad (116)$$

and an explicit relation between I and V_b is obtainable, namely

$$I = I_s(e^{V_b/W_t} - 1) \quad (117)$$

which is the familiar junction rectification equation, where

$$I_s \equiv \frac{eD_n n_p}{w} \quad (118)$$

is the reverse saturation current.

If $Z \gg 1$, the injected electron density is much greater than the equilibrium hole density in the p -region, and the parametric equations reduce to

$$V_b = W_t \log_e \left(\frac{p_p Z^2}{n_p} \right) \quad (119)$$

$$I = \frac{2eD_n p_p Z}{w} \quad (120)$$

$$Q = \frac{ep_p w}{2} Z \quad (121)$$

$$\tau_t = \frac{1}{2} \frac{w^2}{2D_n} \quad (122)$$

Again an explicit relation between I and V is obtainable, being

$$I = I'_s e^{V_b/2W_t} \quad (123)$$

where

$$I'_s \equiv \frac{2eD_n n_i}{w} \quad (124)$$

It should be observed in both eqns. (118) and (124) that w is a function of V_b , since the thickness of the charge-neutral p -region is equal to the total physical length of the p -region minus the transition region thickness, which is a function of V_j . This effect is of considerable importance in transistors, but is not of importance in diodes unless the thickness of the p -region is very small.

The 2-electrode semiconductor diode is a degenerate form of charge-control device in which the input and output terminals are the same, namely the anode. As for the vacuum diode, subscripts 1 and 2 in the charge-control formulae of Section 3 can be omitted when calculating the incremental parameters, and thus (for low-level injection)

$$C \equiv \frac{\partial Q}{\partial V_b} = \frac{Q}{W_t} = \frac{\tau_t I}{W_t} \quad (125)$$

$$G \equiv \frac{\partial I}{\partial V_b} = \frac{(1+n)C}{\tau_t} = \frac{I}{W_t} \quad (126)$$

since $n = 0$ for charge-neutral current flow in a semiconductor. In contrast to the vacuum diode, both the incremental capaci-

tance and conductance increase in direct proportion to the current.

For reverse bias, C is as given by eqn. (94) and $G = 0$.

(6.2) The p - n Junction Transistor

In order to provide a gate mechanism in a semiconductor diode and thus to create a 3-terminal charge-control device, means must be found to control the current of a p - n junction by one electrode and yet to collect the current at an independent electrode.

The problem is in principle simpler than that of introducing a gate into an electron stream between cathode and anode in a vacuum. The minority carrier current of electrons injected into the p -region of a p - n junction is determined by the potential of the p -region with respect to the n -region, and flows through the p -region. Now if the p -region ohmic contact is moved around to the side of the p -region, and another n -region is attached to the large-area face of the p -region, a 3-electrode structure results. The p -region is, in fact, an ideal gate controlling the current flowing from one n -region to the other: it is an equipotential region whose potential is determined by the charge placed in the p -region and yet is permeable to electrons. The conditions in the original p - n junction are almost the same as those already discussed for the p - n junction rectifier; the only difference is that because of the modified geometry the injected electron current arrives at the second n -region instead of at the outside terminal whose voltage is V_b . The only requirement on the second n -region is that it should have a positive potential, so that the second p - n junction is biased in the reverse direction and any electrons which reach the junction are swept across into the n -region. The three electrodes which correspond to the cathode, grid and anode of a vacuum tube are in the semiconductor triode (transistor) the emitter, base and collector, respectively.

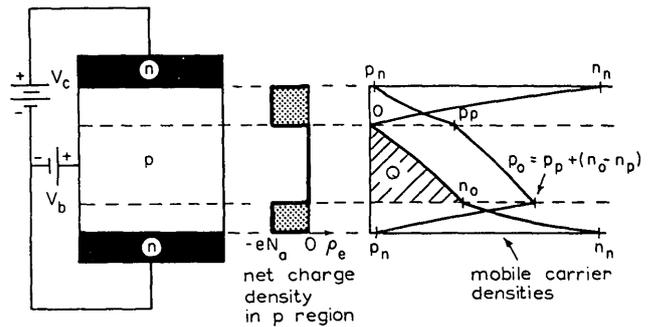


Fig. 12.—Geometry of an n - p - n junction transistor, and electron and hole densities in the p -region (high-level injection).

Fig. 12 shows the n - p - n transistor structure and the electron and hole density distributions in the p -region for forward-biased base-emitter junction and reverse-biased collector-base junction. It is assumed that both the emitter and collector n -regions are of very high conductivity and that in consequence their injection efficiencies are essentially unity. There are charge-depletion regions at both ends of the p -region, and the active base region of thickness w is the charge-neutral part of the p -region between the two charge-depletion regions.

Conditions in the base region of an n - p - n transistor are the same as those in the p -region of a p - n junction diode, except that the boundary value of the electron density at the collector side of the base region is $n \approx 0$ rather than $n = n_p$, because of the reverse bias at the collector. The parametric equations for the base-emitter voltage, V_b , collector-emitter (conventional)

current density, I , total injected electron charge, Q , in the base, and transit time, τ_t , are therefore the same as given in eqns. (104) and (110)–(112), except that the terms in n_p in eqns. (110)–(112) are zero:

$$V_b = W_t \log_e \left[\frac{p_p Z(1 + Z)}{n_p} \right] \quad (127)$$

$$\frac{w}{eD_n p_p} I = 2Z - \log_e(1 + Z) \quad (128)$$

$$Q = \frac{ep_p w}{2} Z \quad (129)$$

$$\tau_t = \frac{w^2/2D_n}{2 - \frac{\log_e(1 + Z)}{Z}} \quad (130)$$

For low-level injection ($Z \ll 1$) these relations reduce to

$$V_b = W_t \log_e(n_0/n_p) \quad (131)$$

$$I = \frac{eD_n n_0}{w} \quad (132)$$

$$Q = \frac{en_0 w}{2} \quad (133)$$

$$\tau_t = w^2/2D_n \quad (134)$$

The basic charge-control relations of Section 3 can now be applied to obtain the incremental parameters, giving

$$C_1 \equiv \frac{\partial Q}{\partial V_b} \Big|_{V_c} = \frac{Q}{W_t} = \frac{\tau_t I}{W_t} \quad (135)$$

$$g_m = \frac{(1 + n)C_1}{\tau_t} = \frac{I}{W_t} \quad (136)$$

$$C_2 = G_2 = G_1 = 0 \quad (137)$$

The incremental equivalent circuit is therefore as shown in Fig. 2, but with C_2 , G_1 and G_2 omitted, as in the case of the ideal vacuum tube. However, just as the grid of a practical vacuum tube falls short of being ideal, so does the base region of a transistor. Departures from ideality in the transistor arise from two sources: first, the unavoidable presence of recombination in the base region (so far neglected) introduces base input conductance G_1 ; second, the dependence of the two transition-region thicknesses on the emitter and collector voltages causes the base thickness w to be likewise dependent, and thus introduces output conductance and capacitance G_2 and C_2 because of the dependence of I and Q on w . These departures from ideality are sufficiently great to require inclusion even in a basic treatment of the transistor.

In discussing the departure from ideality due to the practical grid in a vacuum triode, it was convenient to introduce an arbitrary parameter μ which was subsequently found to be the ratio between the incremental anode and grid voltages at constant anode current. This procedure by-passed the necessity of investigating in detail the geometry of the grid structure. It is convenient to proceed in an analogous manner in attempting to account for base-region recombination in a transistor, since it is difficult to calculate the base recombination current without detailed knowledge of the base geometry, physical properties and surface condition. Let an arbitrary parameter β be introduced, defined as the ratio between the incremental collector and base currents at constant collector voltage, i.e.

$$\beta \equiv \frac{\partial I}{\partial I_b} \Big|_{V_c} \quad (138)$$

Unfortunately, β is not as independent of collector voltage and

current as μ is of anode voltage and current, but sufficiently accurate results for most purposes can be obtained by assuming β to be constant.

The effect of the collector and emitter voltages on the base width can easily be introduced. The transition-region thickness as a function of junction voltage is given in eqn. (93), and the effective base thickness, w , can then be found if the physical thickness of the p -region is known. For calculation of incremental parameters, only the change in the collector transition-region thickness need be considered, since the base-emitter incremental voltage is very small. Moreover, it is more convenient to leave formulae in terms of $\partial w/\partial V_c$, where V_c is the collector reverse-bias voltage, rather than to substitute an explicit expression for this dependence. Although $\partial w/\partial V_c$ is a negative number, it will be assumed positive in order to avoid writing minus signs in the expressions for element values in the incremental equivalent circuit.

With the inclusion of the effects of base recombination and voltage-dependent base width, application of the basic charge-control relations to eqns. (131)–(134) gives, for low-level injection,

$$C_1 \equiv \frac{\partial Q}{\partial V_b} \Big|_{V_c} = \frac{\tau_t I}{W_t} \quad (139)$$

$$G_1 \equiv \frac{\partial I_b}{\partial V_b} \Big|_{V_c} = \frac{\partial I_b}{\partial I} \frac{\partial I}{\partial V_b} \Big|_{V_c} = \frac{1}{\beta} \frac{I}{W_t} \quad (140)$$

$$C_2 \equiv \frac{\partial Q}{\partial V_c} \Big|_{V_b} = \frac{\partial Q}{\partial w} \frac{\partial w}{\partial V_c} \Big|_{V_b} = \frac{\tau_t I}{w} \frac{\partial w}{\partial V_c} \quad (141)$$

$$G_2 = \frac{(1 + n)C_2}{\tau_t} = \frac{I}{w} \frac{\partial w}{\partial V_c} \quad (142)$$

$$g_m = \frac{(1 + n)C_1}{\tau_t} = \frac{I}{W_t} \quad (143)$$

When drawing the incremental equivalent circuit it must be remembered that the five parameters C_1 , G_1 , C_2 , G_2 , g_m have all been defined as short-circuit quantities. Thus it is not immediately known, for example, whether the output conductance G_2 actually exists between collector and emitter, or whether some of this conductance is between collector and base. However, the difficulty can be resolved by recalling that, by assumption, a fraction $1/\beta$ of the collector current flows in the base; hence the output conductance G_2 is divided in such a way that a fraction G_2/β exists between the collector and the base, and a fraction $G_2(1 - 1/\beta)$ exists between the collector and the emitter. The same conclusions apply to the division of the output capacitance C_2 . Fig. 13 shows the intrinsic equivalent circuit embodying the results of the above discussion.

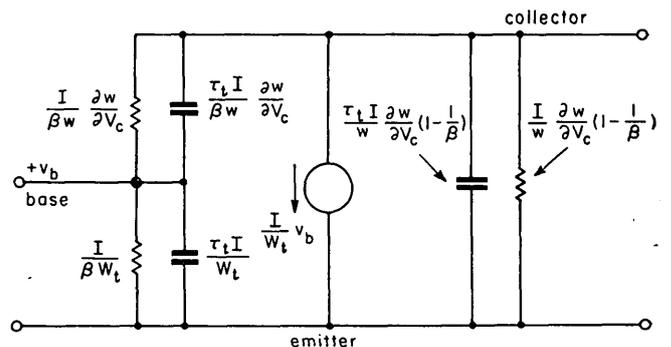


Fig. 13.—Incremental model of a junction transistor, showing ‘parasitic’ elements in addition to the ‘essential’ elements g_m and C_1 (low-level injection).

In addition to the intrinsic elements, the extrinsic elements C_{te} , C_{tc} and r_b should be included in a complete equivalent circuit, where r_b is here the ohmic base resistance. Some simplifications may be made, however: C_{te} is often much smaller than C_1 , and may be omitted; C_2 is small enough to be omitted at most frequencies at which the equivalent circuit is valid; $G_2(1 - 1/\beta)$ may be written simply as G_2 , since in any useful transistor β is much greater than unity. Thus the complete equivalent circuit including intrinsic and extrinsic elements of importance is as shown in Fig. 14, and is recognized to be identical with the hybrid- π model introduced by Giacoletto.³

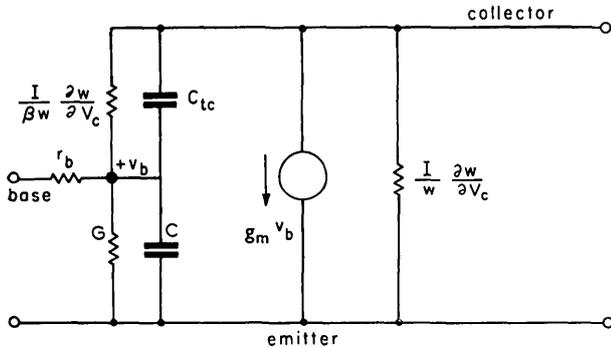


Fig. 14.—Simplified incremental model of a junction transistor, with the addition of the 'extrinsic' elements C_{tc} and r_b .
 For low-level injection, $G = I/\beta W_t$, $C = I\tau_{t0}/W_t$, $g_m = I/W_t$.
 For high-level injection, $G = I/2\beta W_t$, $C = I\tau_{t0}/4W_t$, $g_m = I/2W_t$.

The incremental model discussed above is valid only for low-level injection. To obtain a model valid for arbitrary injection levels the general expressions of eqns. (127)–(130) must be used in deriving the incremental parameters. The carrier densities in the charge-neutral part of the base region are shown in Fig. 12. The results are

$$C_1 = \frac{\tau_t I}{W_t} \frac{1 + Z}{1 + 2Z} \quad (144)$$

$$G_1 = \frac{I}{\beta W_t} \frac{1}{2 - \frac{\log_e(1 + Z)}{Z}} \quad (145)$$

$$C_2 = \frac{\tau_t I}{w} \frac{\partial w}{\partial V_c} \quad (146)$$

$$G_2 = \frac{I}{w} \frac{\partial w}{\partial V_c} \quad (147)$$

$$g_m = \frac{I}{W_t} \frac{1}{2 - \frac{\log_e(1 + Z)}{Z}} \quad (148)$$

where $\tau_t = \frac{\tau_{t0}}{2 - \frac{\log_e(1 + Z)}{Z}}$, $\tau_{t0} \equiv w^2/2D_n$ (149)

The same arguments used in setting up the complete incremental model for low-level injection are also applicable for arbitrary injection level. The results for high-level injection ($Z \gg 1$) are shown in Fig. 14.

(6.2.1) Junction-Transistor Saturation Currents.

Incremental equivalent circuits for a junction transistor biased in the normal amplifying manner are developed in the preceding Section. In order to establish a given operating point, suitable bias circuits must be employed to minimize variations due to changes in the d.c. characteristics of the transistor.

The two transistor parameters of most importance in establishing the d.c. operating point are the collector saturation current and the base-emitter voltage. The collector saturation current is not controlled by the input voltage and, when the collector is reverse-biased, is present even when the input voltage is zero. Under these conditions the electron density at the emitter side of the charge-neutral base region of an $n-p-n$ transistor is n_p , and that at the collector side is practically zero. Under the usual assumption of very small recombination rate in the base region, the electron density is nearly linear within the charge-neutral base region and a diffusion current of magnitude $eD_n n_p/w$ exists at the collector. This quantity has previously been defined as I_s in eqn. (118). Because the electron density in the base is not equal to its equilibrium value, recombination and generation will not balance and a base current will therefore be present. Let the base current be a fraction, $1/\beta$, of the collector current I_s , where it must be borne in mind that the value of β at such low currents is frequently much smaller than that defined for high operating currents by eqn. (138). Under these conditions of short-circuited base-emitter and reverse-biased collector, the electron density and current distributions in the base of an $n-p-n$ transistor are as shown in Figs. 15(a) and (b).

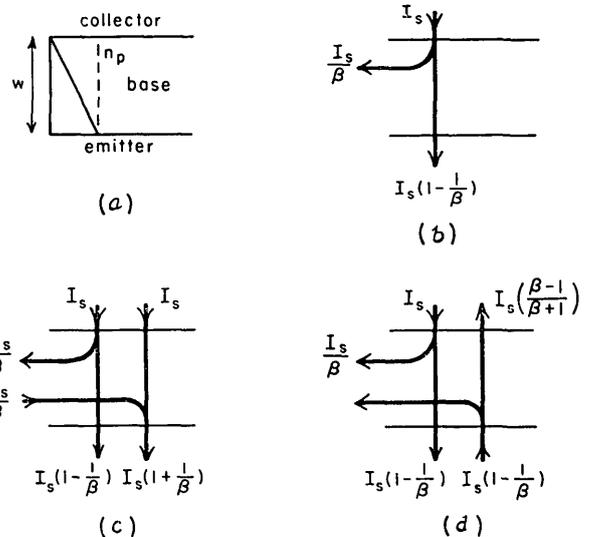


Fig. 15.—Collector saturation currents in an $n-p-n$ transistor under various conditions.

The collector junction is reverse biased in all cases. The electron distribution in the base for $V_b = 0$ is shown in (a). Current distributions for $V_b = 0$, $I_b = 0$, and $I_e = 0$ are shown in (b), (c), and (d) respectively.

Now let the short-circuit be removed from the base-emitter and the base left floating. If β is assumed to remain constant, then superposition may be used to determine the new collector current. Conditions are as shown in Fig. 15(c), in which an additional base current I_s/β is required to cancel that present when $V_b = 0$. It then follows that the collector current with open base, defined as I_{ce0} , is

$$I|_{I_b=0} \equiv I_{ce0} = 2I_s \quad (150)$$

If the emitter instead of the base is left open, the corresponding derivation of the new collector current requires an additional emitter current $I_s(1 - 1/\beta)$ to cancel that present when $V_b = 0$, as shown in Fig. 15(d). It then follows that the collector current with open base, usually known as I_{c0} , is

$$I|_{I_b=0} \equiv I_{c0} = I_s - I_s \left(\frac{\beta - 1}{\beta + 1} \right) = \frac{2I_s}{1 + \beta} \quad (151)$$

From eqns. (150) and (151) it is seen that $I_{ce0} = (1 + \beta)I_{c0}$, a familiar relation. However, the derivation indicates that I_{ce0} is a more fundamental quantity than I_{c0} , in that the saturation current I_{ce0} is characteristic of the ideal transistor and its value is independent of any non-ideality. In contrast, I_{c0} is strongly dependent upon β , which has been shown to be a measure of the departure of the practical transistor from ideal. The basic relation between the direct currents is therefore

$$I = \beta I_b + I_{ce0} \quad \dots \quad (152)$$

in preference to the more usual

$$I = \frac{\beta}{1 + \beta} I_e + I_{c0} \quad \dots \quad (153)$$

The temperature dependence of I_{ce0} —of special importance in the design of bias networks—is essentially that of n_p . The equilibrium minority-carrier density, n_p , in the base increases rapidly with temperature, with a corresponding increase in I_{ce0} . The dependence is usually given as a percentage increase, and approximately

$$\frac{dI_{ce0}}{I_{ce0}} = \left\{ \begin{array}{l} 9.3\% \text{ per deg C for germanium at } 27^\circ \text{ C} \\ 14\% \text{ per deg C for silicon at } 27^\circ \text{ C} \end{array} \right\} \quad (154)$$

The base-emitter direct voltage for arbitrary injection level is given in eqn. (127) in terms of the injection-level parameter Z . The collector current I is given in terms of Z in eqn. (128). If the current is held constant with change in temperature, it is seen that Z is essentially constant (unless the intrinsic temperature range is approached, in which case p_p ceases to be constant) and the temperature dependence of V_b is contained in W_i and n_p only. Since the equilibrium minority-carrier density in the base may be expressed in the form

$$n_p = N e^{-W_0/W_i} \quad \dots \quad (155)$$

where the temperature dependence of the constant N is negligible compared with that of the exponential term, the base-emitter voltage V_b may be written

$$V_b = W_i \log_e \left[\frac{p_p Z(1 + Z)}{N} e^{W_0/W_i} \right] \quad \dots \quad (156)$$

Hence
$$\frac{dV_b}{dT} = \frac{k}{e} \log_e \left[\frac{p_p Z(1 + Z)}{N} \right] \quad \dots \quad (157)$$

$$= - \left(\frac{W_s - V_b}{T} \right) \quad \dots \quad (158)$$

Although this last result appears to show that dV_b/dT is a function of temperature, this is not true, since V_b and T are not independent. Eqn. (157) shows that dV_b/dT is indeed independent of temperature (below the intrinsic temperature range) at any injection level, although it does depend on injection level. Approximate numerical values are

$$\frac{dV_b}{dT} = \left\{ \begin{array}{l} -1.4 \text{ mV per deg C for germanium} \\ -2.0 \text{ mV per deg C for silicon} \end{array} \right\} \quad (159)$$

(7) CONCLUSIONS

An attempt to develop the theory of vacuum tubes and transistors as examples of charge-controlled devices has been made. The emphasis has been on the basic similarity of these devices, and in order to illuminate the underlying unity of the approach, only the salient features of the devices have been discussed. In spite of the simplicity of the analysis, many results have been

derived which are obtained only with considerably greater complexity by the classical approaches. Although devices more elaborate than the vacuum triode and the diffusion-limited junction transistor have not been discussed, the methods are applicable to multi-electrode vacuum tubes, gas tubes, drift transistors and to the new class of solid-state devices based on space-charge-limited current flow.²⁰ Much remains to be done in the development of a fully-integrated treatment of vacuum and solid-state devices. It is felt that the approach here described is a powerful yet simple tool for understanding a rapidly growing class of electronic devices.

(8) REFERENCES

- (1) JOHNSON, E. O., and ROSE, A.: 'Simple General Analysis of Amplifier Devices with Emitter, Control, and Collector Functions', *Proceedings of the Institute of Radio Engineers*, 1959, **47**, p. 407.
- (2) BEAUFOY, R., and SPARKES, J. J.: 'The Junction Transistor as a Charge-Controlled Device', *A.T.E. Journal*, 1957, **13**, p. 310.
- (3) GIACOLETTO, L. J.: 'Study of $p-n-p$ Alloy Junction Transistor from D.C. through Medium Frequencies', *RCA Review*, 1954, **15**, p. 506.
- (4) WEBSTER, W. M.: 'A Comparison of Analogous Semiconductor and Gaseous Electronics Devices', from 'Advances in Electronics and Electron Physics' (Academic Press, 1954), Volume 6, p. 257.
- (5) SHOCKLEY, W., and PRIM, R. C.: 'Space-Charge-Limited Emission in Semiconductors', *Physical Review*, 1953, **90**, p. 753.
- (6) PARKER, P.: 'Electronics' (Arnold, 1950), pp. 650-657 and 696.
- (7) JEANS, J.: 'An Introduction to the Kinetic Theory of Gases' (Cambridge, 1940), p. 52.
- (8) SHOCKLEY, W.: 'Electrons and Holes in Semiconductors' (Van Nostrand, 1950), pp. 191-195 and 200-204.
- (9) MILLMAN, J., and SEELY, S.: 'Electronics' (McGraw-Hill, 1941), First edition, p. 136.
- (10) *Ibid.*, p. 244.
- (11) *Ibid.*, p. 152.
- (12) MOULLIN, E. B.: 'On the Amplification Factor of the Triode', *Proceedings I.E.E.*, Monograph No. 211 R, November, 1956 (**105 C**, p. 196).
- (13) THOMPSON, B. J.: 'Space-Current Flow in Vacuum-Tube Structures', *Proceedings of the Institute of Radio Engineers*, 1943, **31**, p. 485.
- (14) DOW, W. G.: 'Fundamentals of Engineering Electronics' (Wiley, 1952), Second Edition.
- (15) SCHOTTKY, W.: 'Über Hochvakuumverstärker. III Teil-Mehrgitterröhren', *Archiv für Elektrotechnik*, 1919, **8**, p. 299.
- (16) SPENKE, E.: 'Electronic Semiconductors' (McGraw-Hill, 1958).
- (17) VAN DER ZIEL, A.: 'Solid-State Physical Electronics' (Prentice-Hall, 1957).
- (18) MIDDLEBROOK, R. D.: 'An Introduction to Junction Transistor Theory' (Wiley, 1957), p. 104.
- (19) WEBSTER, W. M.: 'On the Variation of Junction Transistor Current Amplification Factor with Emitter Current', *Proceedings of the Institute of Radio Engineers*, 1954, **42**, p. 914.
- (20) WRIGHT, G. T.: 'Some Properties and Applications of Space-Charge-Limited Currents in Insulating Crystals' (see page 915).