Transistor Switching Analysis*

Carver A. Mead

Abstract—With the widespread application of junction transistors in switching applications, the need for a general method of analysis useful in the region of collector voltage saturation has become apparent. Linear equivalent circuits using lumped elements have long been used for small signal calculations of normally-biased transistors, but a comparable method for saturated transistors has been lacking. Recently, Linvill [3] proposed the method of lumped models which allow the analysis of complex switching problems with the ease of linear circuit calculations. The method is shown to be equivalent to a well-known linear equivalent circuit under normal bias conditions. Examples of the application of the method and the use of approximations are drawn from practical circuit problems. Emphasis is placed upon the understanding of the physical phenomena involved, a necessary prerequisite to intelligent circuit design.

1 Introduction

With the first analysis of a junction transistor triode, it was recognized that such a device was capable of *symmetrical* operation; that is, either the "emitter" junction or "collector" junction could act as a source of minority carriers in the base region. Thus, modes of operation are available in a transistor which have never existed in the vacuum tube. For example, a saturated transistor (both emitter and collector forward biased) will carry signals well in both directions, while if both junctions are reverse biased, essentially no signal is allowed to pass in either direction. The inherently low voltage drop across a saturated transistor makes possible the control of very high powers with low dissipations. For these reasons, transistors find switching service a most important and useful application.

The first detailed analysis of the large signal properties of transistors was done by Ebers and Moll [1]. Later, Linvill [3] proposed a technique by which the same results may be obtained, but which has the advantage that a linear model is used and physical insight into the behavior of the device is more readily gained. The purpose of the present article is to extend this method to the general treatment of diodes and transistors in practical circuit applications, to present results in special cases of importance, and to illustrate applications of the analysis in sufficient detail to be generally useful to the design engineer.

1.1 Requirements of the Analytical Method

Let us examine the requirements on a method of analysis to be used in problems of this nature. Clearly, what is needed is a model, similar to an equivalent circuit such as used for small signal work, yet appropriate for all conditions encountered in transistor operation; i.e., either junction may be forward or reverse biased, currents may be either small or large, *a-c, d-c,* or both.

*Re-typeset from original material by Donna Fox, October 2016. This paper was originally published in *Semiconductor Products,* Vol. 3 (9), pp. 43-47, 1960 (Part 1); Vol. 3 (10), pp. 38-42, 1960 (Part 2); and, Vol. 3 (11), pp. 28-32, 1960 (Part 3).

Together with a model, the method must include a procedure for evaluating elements in the model and for making approximations where an exact analysis would be too cumbersome. Furthermore, in order to be of general utility, the method of analysis should possess the following qualities:

a) The principal variables should be related in a linear manner so linear circuit theory may be used.

b) Non-linearities in the system should be easily and accurately approximated by a simple piecewise linear idealization.

c) Elements in the model should be readily obtainable in terms of simple, easily-measured device parameters.

d) Variables in the model should possess physical significance, and results of the analysis should enhance one's physical insight into the problem.

e) The model should reduce to familiar form for special cases, e.g., normal bias, small signal.

The variables necessary to solve normal semiconductor problems are:

a) Junction voltage

b) Junction current

c) Minority carrier density

provided we deal only with devices in which the diffusion current predominates. Transistors of this type are typically used in switching service where the method is most generally useful.

1.2 Review of Basic Processes

Before launching into the details of analysis, let us briefly review the basic processes which occur within a semiconductor. If we limit our discussion to one-dimensional diffusion flow, a complete description of the motion of minority carriers within the material consists of:

1. The continuity equation for minority carriers

$$\frac{\partial N}{\partial t} = \frac{N_0 - N}{\tau} + D\frac{\partial^2 N}{\partial x^2} \tag{1}$$

where

N is the density (number per unit volume) of minority carriers as a function of x and t

N_0 is the density of minority carriers at thermal equilibrium

τ is the "lifetime" of minority carriers

D is the diffusion constant of minority carriers

x is the distance through the semiconductor

2. The condition that any macroscopic volume element of the material be electrically neutral.

1.3 Introduction to Lumped Models

The continuity equation is a partial differential equation, involving both time and space derivatives and its solutions are in general both difficult and messy, resulting in carrier densities at all points as a function of time.

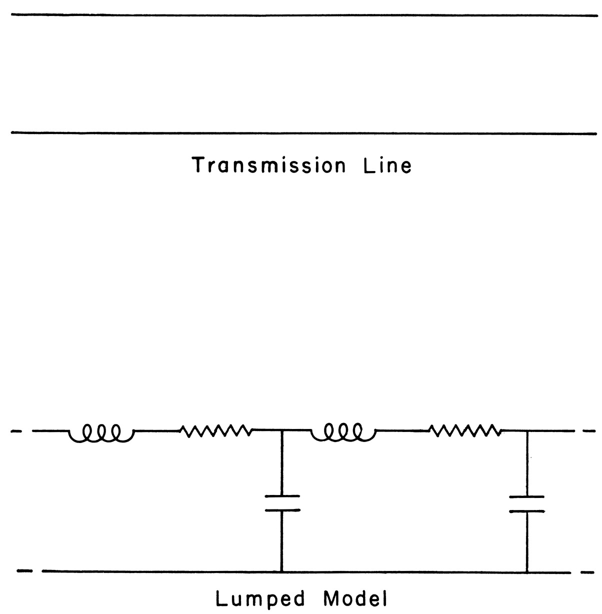


Figure 1: Transmission line and equivalent lumped element representation.

One is reminded of the transmission line problem where again a partial differential equation must be solved, and the results are voltage (or current) at any point on the line as a function of time. Pursuing the analog still further, the analysis of a transmission line is greatly simplified if we are content to find *approximate* voltages and currents at *certain specified* points along the line, instead of *exact* voltages and currents at *all* points along the line. Since normally, only the ends of the line are of interest, such a procedure seems highly desirable. To this end, we approximate the line by a ladder network as shown in Fig. 1.

The line possesses a series inductance and resistance per unit length, and a shunt capacitance per unit length. These distributed parameters are represented in the ladder network by the *lumped elements* L, R and C. Thus, we have created a *lumped model* as an approximation to the real transmission line. The voltage at any node, or the current through any element, may be obtained by standard circuit analysis techniques. Hence, we have transformed a problem in partial differential equations into a problem in simple circuit theory.

Each section of the lumped model corresponds to a given length of transmission line. As the number of sections is increased, the length of line to which each section corresponds is decreased, and the accuracy of the approximation is improved. In the limit, as the number of sections becomes infinite, each section represents an infinitesimal length of time, and we are again faced with the solution of a partial differential equation. Perhaps the most significant feature of the lumped model is that the approximations have been made before any equations were written, and each element in the model has definite physical significance.

1.4 Semiconductor Lumped Model

In order to simplify our expressions for minority carrier flow, it is convenient to write the continuity equation in terms of a new variable, the *excess density*, $N - N_0$.

$$\frac{\partial(N - N_0)}{\partial t} = -\frac{(N - N_0)}{\tau} + D\frac{\partial^2(N - N_0)}{\partial x^2} \tag{2}$$

For a p-type semiconductor: $N = n, N_0 = n_p$

Therefore,

$$\frac{\partial \eta}{\partial t} = \frac{\eta}{\tau} + D\frac{\partial^2 \eta}{\partial x^2} \qquad \eta = n - n_p \tag{3}$$

For an n-type semiconductor: $N = p, N_0 = n_n$

Therefore,

$$\frac{\partial \rho}{\partial t} = -\frac{\rho}{\tau} + D\frac{\partial^2 \rho}{\partial x^2} \qquad \rho = p - p_n \tag{4}$$

Now let us examine the physical significance of each term in the continuity equation. Consider a long bar of n-type semiconductor with unit cross-sectional area containing volume elements 1 and 2.

Suppose within the bar there exists a distribution of excess holes $\rho(x)$ as shown in Fig. 2, which is a function of x, but independent of y and z. The volume elements will possess average excess densities $\rho 1$ and $\rho 2$ respectively.

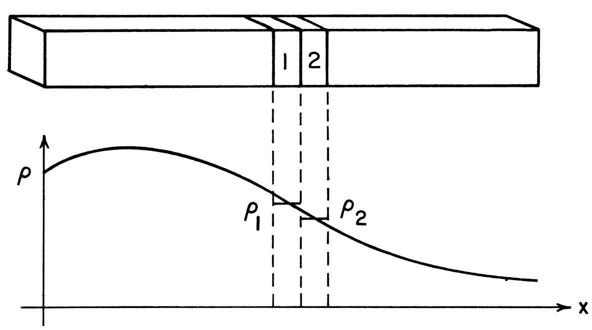


Figure 2: Semiconductor bar with excess minority carrier distribution.

Storage. Due to the charge neutrality requirement, an excess density of minority carriers implies an equal excess density of majority carriers. Thus, a *change in minority excess density with time produces a majority carrier current into the volume element.* Such a change in density with time is represented by the first term in the continuity equation, and the resulting current is the *rate of change of stored charge* in the volume element.

Recombination. According to the simple linear recombination law, the recombination rate is proportional to the excess density. Again, the charge neutrality condition requires that if the minority carrier density remains fixed, for each recombination a new majority carrier must enter the volume element. However, in order for the minority carrier density to remain fixed, a new minority carrier must also enter the volume element. Thus, since the carriers possess opposite charges, the net effect of recombination is to bring minority carrier current into the volume element and force an equal majority current out of the volume element. This effect is represented by the second term in the continuity equation.

Diffusion. If the slope of the carrier excess density curve is greater at the left boundary of the volume element than at the right boundary, there will be a net minority carrier diffusion current into the volume element. This effect is represented by the last term in the continuity equation. Since the diffusion current is proportional to the gradient of the excess density, the current flowing from element 1 into element 2 will be approximately proportional to $(\rho_1 - \rho_2)$.

We may state the continuity equation, then, as follows:

The sum of currents flowing out of a given volume element due to

a) change in stored charge

b) recombination of carriers

c) minority carrier diffusion is equal to zero.

When so stated, it is clear that the continuity equation is merely Kirchhoff's law for a continuous system. However, we have seen that such a system may be approximated by a lumped model in which the network node equations replace the continuity equation. We may therefore construct such a model for our bar of semiconductor. Lumped models of this type were first proposed by Linvill [3].

Lumped Model Elements. The elements required for such a lumped model are clearly of a different nature than those of electric circuit theory. The variables are excess density and current rather than voltage and current. It should always be borne in mind that the semiconductor lumped model is analogous, not identical, to its corresponding electric circuit. However, the same methods are available for the solution of either type of problem.

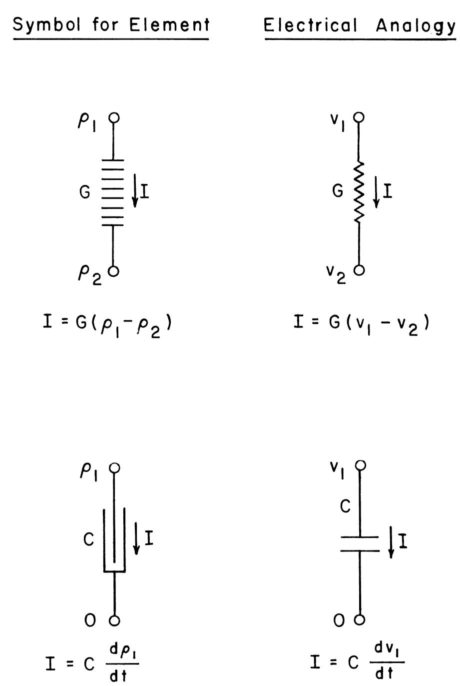


Table I

The types of elements required may be summarized as in Table I. Clearly, the excess densities are the direct analogies of voltages in an electrical circuit.

The lumped model for our bar of semiconductor may be constructed in exactly the same manner as for the transmission line, as shown in Fig. 3. Here, ρ_1 and ρ_2 are the average excess hole densities in volume elements 1 and 2. G_1 and C_1 represent recombination and storage in volume element 1, while G_2 and C_2 represent recombination and storage in volume element 2. G_{12} represents the diffusion of holes from volume element 1 into volume element 2. It will be noticed that the lines are not wires, but are defined as regions of constant excess density. The bottom line is defined as the zero excess density, and current flowing in this line is majority carrier current. This convention follows from the definitions of the elements and the effects which they represent. Although the lumped model elements are linear, it should be borne in mind that the excess density ρ is constrained to be greater than $-p_n$ at all times, since the actual density cannot become negative. This restriction will introduce nonlinearities later when specific problems are considered.

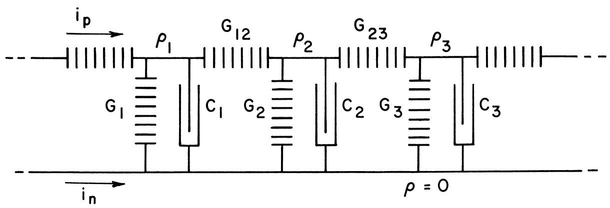


Figure 3: Lumped model of semiconductor bar.

As a check on the lumped model, let us assume a uniform excess minority carrier density in the semiconductor bar. Under these conditions, the continuity equation becomes

$$\frac{\partial \rho}{\partial t} = -\frac{\rho}{\tau} \tag{5}$$

which has the solution

$$\rho = \rho_0 e^{-t/\tau} \tag{6}$$

From the lumped model

$$\rho_1 = \rho_2 = \rho_3 \text{ etc.} \tag{7}$$

Thus, no current flows in the diffusion conductances, and we may write Kirchhoff's law as, for example

$$\rho_2 G_2 + C_2 \frac{\partial \rho_2}{\partial t} = 0 \tag{8}$$

or

$$\frac{\partial \rho}{\partial t} = -\frac{G_2}{C_2}\rho_2 \tag{9}$$

which has the solution

$$\rho_2 = \rho_0 \, e^{-\frac{G_2}{C_2}t} \tag{10}$$

and similarly for the other junctions. Thus, the lifetime τ may be identified with the time constants of the lumped model parallel elements

$$\frac{1}{\tau} = \frac{G_1}{C_1} = \frac{G_2}{C_2} = \frac{G_3}{C_3} \text{ etc.} \tag{11}$$

It is now clear that the basic diffusion process within the semiconductor is extremely simple, analogous to an RC ladder network where all capacitors are connected to ground, and all shunt elements have time constant τ. No inductive elements are present, since the diffusion process cannot support a propagating wave. Our lumped approximation is complete, and we are in a position to solve any problem where the initial conditions or boundary conditions on the excess density are known, by the application of simple circuit theory.

Boundary Conditions, the *p-n* Junction. In a semiconductor device such as the transistor, the boundary conditions on minority carrier excess density are generally determined by the voltage across certain *p-n* junctions in the structure. Depending upon the type of semiconductor material, temperature, operating current density and other factors, the value of excess density determined by a junction with a certain applied potential may vary. However, in most instances, the recombination within the junction is small, and the following expressions are sufficiently accurate for practical application:

$$\begin{aligned} \rho_j &= p_n(e^{qv/kT} - 1) \\ \eta_j &= n_p(e^{qv/kT} - 1) \end{aligned} \tag{12}$$

where

ρ_j is the excess hold density at the n side of the junction

η_j is the excess electron density at the p side of the junction

k is Boltzmann's constant

T is the absolute temperature

v is the voltage across the junction, being taken as positive when the junction is forward biased

In this approximation, the hole and electron currents entering the junction from one region emerge undiminished at the opposite side of the junction. Thus, the junction provides an excess density essentially independent of current, or acts as an *excess density generator* controlled by the junction voltage. The analogy in electric circuit theory would be a *voltage generator* controlled by some other variable in the system, as frequently used in small signal-equivalent circuits for active devices. The symbol for such

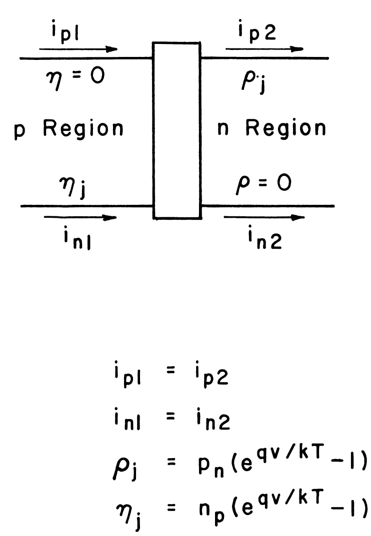


Figure 4: Model of a *p-n* junction.

a junction is shown in Fig. 4. Here again, the top lines carry hole current, and the bottom lines carry electron current.

1.5 *P-N* Diode

A *p-n* junction diode consists not only of the *p-n* junction itself, but also of the adjoining p and n regions. Thus, the diode model must include lumped element network sections for both regions, as well as the junction "excess density generator." The number of sections used for each region is determined by the accuracy required, and the fortitude of the analyst. An appropriate compromise is usually obtained at one or two sections per region. A two-section model is shown in Fig. 5(a). As before, the junction boundary densities ρ_1 and η_1 are determined by the junction voltage through the exponential junction law.

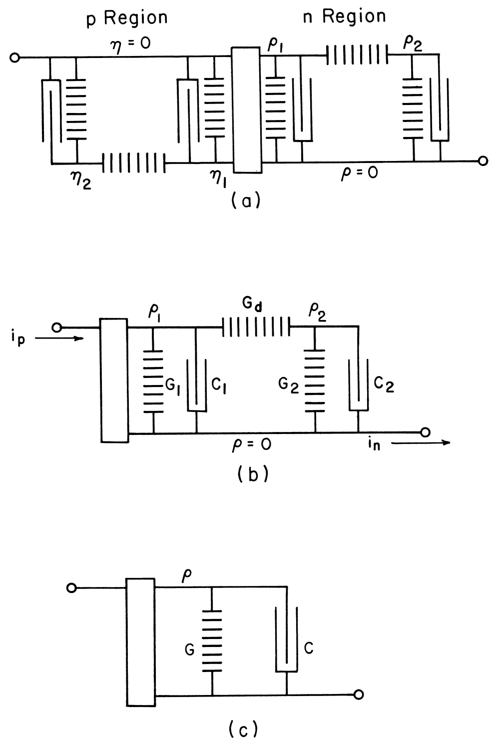


Figure 5: Development of diode model.

In cases where the conductivity of one region is much greater than that of the other, carrier injection into the high conductivity region is small and may be neglected. For example, if the p region is of very high conductivity compared with the n region, n_p is very small compared to p_n (since $p_p n_p = n_i{}^2$, a constant for given temperature). Thus, η_1 will be very small compared to ρ_1, and only hole current need be considered. Under these conditions, considerable simplification in the lumped model is possible, i.e., the elements corresponding to the p region may be deleted, as shown in Fig. 5(b).

For many practical applications, sufficient accuracy is obtained by the use of only one lumped section, as shown in Fig. 5(c). In this model, the capacitance C represents minority carriers stored near the junction, and the conductance G represents the combined effect of recombination near the junction and diffusion, away from the junction and subsequent recombination. As will be seen, the great virtue of the single-section model is the ease with which the element values are determined.

Diode Small Signal Response. If the diode is forward biased and a small *a-c* signal $v_1\ e^{jwt}$ is superimposed on the *d-c* bias voltage v_0, we may write the injected excess density

$$\rho = p_n(e^{qv/kT} - 1) \tag{13}$$

where

$$v = v_0 + v_1\ e^{jwt}. \tag{14}$$

Now, if $v_1 \ll kT/q$, we may expand the exponential and retain only the first two terms

$$\rho \approx p_n \left[e^{qv_0/kT} \left(1 + \frac{q}{kT} v_1\ e^{jwt} \right) - 1 \right] \tag{15}$$

Thus, the *a-c* component of excess density is

$$\rho_1\ e^{jwt} \approx \frac{q}{kT} v_1 p_n\ e^{qv_0/kT}\ e^{jwt} \tag{16}$$

if the *d-c* current through the diode is much greater than its saturation current, the excess density will be much greater than p_n, and the exponential term in the expression for excess density will be large compared to unity. Thus, the *d-c* component of excess density becomes

$$\rho_0 \approx p_n\ e^{qv_0/kT} \tag{17}$$

Combining these two expressions, the amplitude of the *a-c* excess density is

$$\rho_1 \approx \rho_o \frac{q}{kT} v_1 \tag{18}$$

From the lumped model of Fig. 5(c), the current through the diode is composed of a *d-c* bias component

$$I = \rho_o G \tag{19}$$

and an *a-c* component

$$i_1 = \rho_1 (G + j\omega C) \tag{20}$$

But, from Eq. 17 and Eq. 18

$$i_1 = \frac{qI}{kT} \left(1 + j\omega \frac{C}{G} \right) v_1 \tag{21}$$

Thus, the time constant C/G may be determined by one simple small signal measurement on the diode, i.e., the impedance rolloff frequency

$$\omega_h = \frac{G}{C} \tag{22}$$

To determine the magnitude of G, we may measure the *d-c* junction voltage required to produce a certain current. If the current is much larger than the saturation current of the diode

$$I = \rho G \approx p_n\ G\ e^{qv/kT} \tag{23}$$

or

$$p_n G = I\ e^{-qv/kT} \tag{24}$$

It will be noticed that we are not able to determine either p_n or G individually from external measurements on the diode, but only their product. However, for this reason we need not know the individual values if we are interested only in voltages and currents, since expressions for these external variables involve only

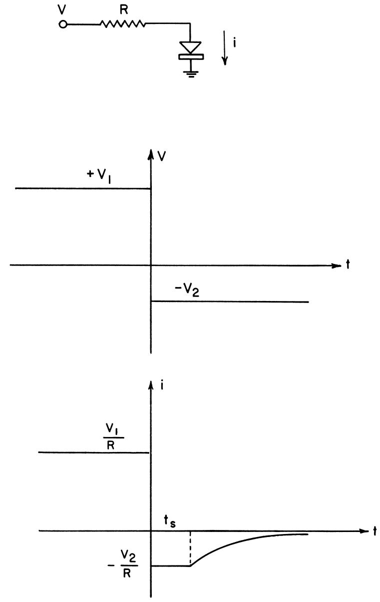


Figure 6: Recovery of *p-n* diode from step input.

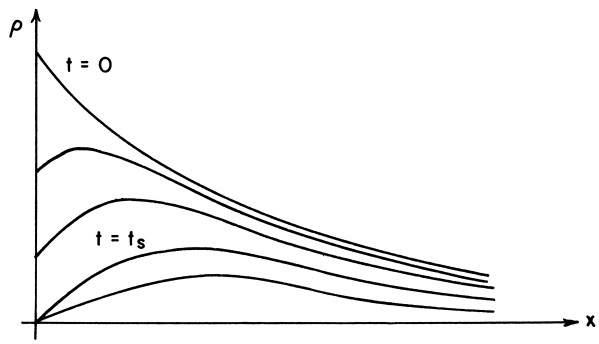


Figure 7: Excess density distribution in *p-n* diode during recovery.

the product. Hence, our two simple measurements on the diode, small signal impedance rolloff frequency, and *d-c* voltage drop for a given *d-c* current, serve to completely determine the lumped model which is now applicable to all operating conditions.

Diode Recovery. Suppose we connect a junction diode in the circuit shown in Fig. 6.

If the applied voltage V remains at $+V_1$ for a sufficient time, the current i will reach a steady-state value very nearly V_1/R, assuming the forward voltage drop across the diode is small compared with circuit voltages. If now the applied voltage is abruptly changed to $-V_2$, the current is observed to assume a nearly constant value $-V_2/R$ for a *storage time* t_s after which it decays rapidly to its small steady-state reverse value i_s. The explanation for this action is as follows [5]:

At $t = 0$, the excess density of minority carriers near the junction is as shown by the top curve of Fig. 7. At a slightly later time, the applied voltage has reversed, but the *stored carriers* have not had time to recombine or diffuse away. The rate at which they may cross the junction is limited to a maximum reverse current of V_2/R, since a higher value would appreciably forward bias the junction and hence be self-annihilating. However, since the excess density of minority carriers at the junction is greater than zero, the *junction must remain slightly forward biased,* and the voltage across the junction remains small. As long as the junction remains forward biased, the current is determined by the external circuit. In this case, the current during the storage period is very nearly $-V_2/R$. This condition determines the slope of minority carrier density at the junction, which is proportional to the junction current. At the end of the storage time t_s, the junction excess density has just reached zero and approaches $-p_n$ as the junction becomes reverse biased. Since the junction excess density is essentially constant (nearly zero) for any reverse-bias voltage, the current is no longer affected by the junction voltage, but is determined only by the conditions within the semiconductor. Therefore, the *decay time* is characterized by the minority carrier lifetime.

Diode Storage Time Calculation. We may obtain a good estimate of the recovery time from a simple calculation of using the one-section lumped model of Fig. 5(c). Although the results are not nearly as accurate as those which will shortly be obtained for the transistor, they are significant, and the method used is typical of all lumped-model calculations.

(a) Steady State

Since the diode is forward biased, the voltage across it is small compared with V_1, and we may assume

$$\frac{V_1}{R} \approx I = G\rho_{ss} \tag{25}$$

The voltage across the diode may be found by using the value of ρ_{ss} obtained from this equation and the exponential junction law.

(b) Storage Period

When the applied voltage changes sign, the charge on C cannot change instantaneously, and hence ρ must be continuous. As long as ρ is greater than zero, the diode is forward biased and the voltage across it is small. Hence, the current is very nearly $-V_2/R$. To determine the storage time t_s, we find ρ as a function of time, assuming the current remains at $-V_2/R$. Then, when ρ becomes very nearly zero, the junction voltage may assume any negative value, and the current approaches zero. To obtain the complete solution for ρ as a function of time, we superpose the steady-state value upon the response to a negative current step of magnitude $(V_1 + V_2)/R$, with the result

$$\rho G = \frac{V_1}{R} - \frac{V_1 + V_2}{R}(1 - e^{-\omega}h^t) \tag{26}$$

At $t = t_s$, the excess density has just reached zero.

$$t_s = \frac{1}{\omega_h} Ln\left(\frac{V_1 + V_2}{V_2}\right) \tag{27}$$

Since there is no longer any charge on C, the current will abruptly stop at $t = t_s$, according to the lumped model. In reality, we know it dies away smoothly. Hence, in this particular case, the lumped model has predicted zero decay time, which may be explained as follows: Since we have used only one lumped section, we have not included the effect of carriers at some distance from the junction, and it is just these carriers which diffuse back to the junction and cause the exponential-type decay. With any finite number of lumped sections, the calculated current will exhibit a discontinuity since ρ at the junction approaches zero with a finite $d\rho/dt$, and causes current through the capacitor nearest the junction which must abruptly cease as ρ is clamped at $-p_n$. This lack of accuracy was introduced because of our attempt to use one lumped section to approximate too large a region of the semiconductor. The problem is especially bad in fast diodes since the lifetime is made very short, and consequently the region to be represented by the lumped model is normally many diffusion lengths. On the other hand, in a transistor, the base region is only a very small fraction of a diffusion length, and the accuracy obtained is very much better. The justification for using the single-section diode model is the ease of determining the element values. In practical design work, the choice is usually not between a simple or elegant analysis, but rather between a simple analysis, or none at all.

1.6 Transistor Lumped Model

If the transistor emitter and collector conductivities are high compared to the base conductivity, we may neglect any carrier injection into the emitter and collector, and consider only minority carriers in the base region.

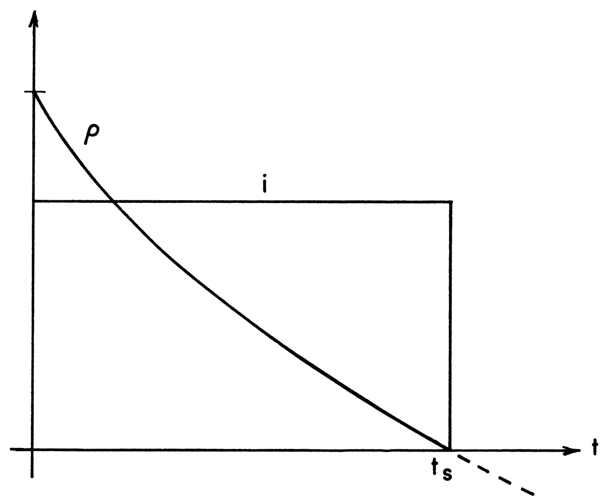


Figure 8: Recovery characteristic predicted by simple single-section model.

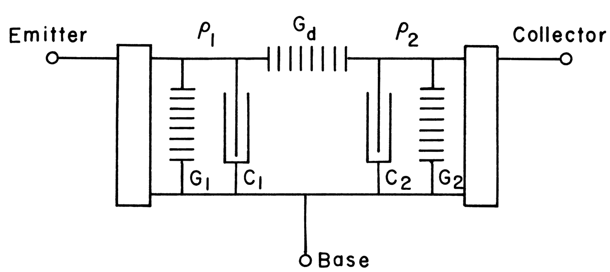


Figure 9: Complete two-section lumped model of junction transistor with high-conductivity collector and emitter regions.

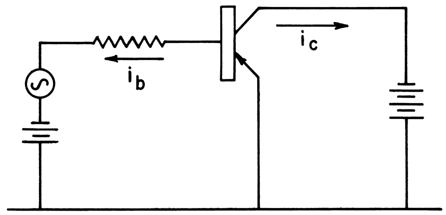
A two-section model results as shown in Fig. 9. In this figure, ρ_1 represents the excess density near the emitter, given by

$$\rho_1 = p_n(e^{qv_{eb}/kT} - 1) \tag{28}$$

where v_{eb} is the emitter-base junction voltage; C_1 and G_1 represent storage and recombination near the emitter, G_d represents diffusion from emitter to collector, ρ_2 is the excess density near the collector, given by

$$\rho_2 = p_n(e^{qv{}_{c}{}^{b/kT}} - 1) \tag{29}$$

where v_{cb} is the collector-base junction voltage; C_2 and G_2 represent storage and recombination near the collector. Both voltages are taken positive when the junction is forward biased.



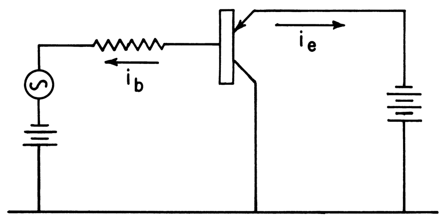


Figure 10: *p-n-p* transistor in (a) normal and (b) inverse connection.

When the transistor is normally biased, as shown in Fig. 10(a), $\rho_2 = -p_n =$ constant, and thus no *a-c* current flows through G_2 or C_2. Therefore, the *a-c* collector current

$$i_c = \rho_1 G_d \tag{30}$$

and the *a-c* base current

$$i_b = \rho_1 (G_1 + sC_1) \tag{31}$$

using the Laplace transform notation. Hence

$$\frac{i_c}{i_b} = \frac{G_d}{G_1 + sC_1} = \frac{G_d/G_1}{1 + s\frac{C_1}{G_1}} = \frac{\beta}{1 + \frac{s}{\omega_\beta}} \tag{32}$$

where β is the low frequency, short circuit, common emitter, current gain and ω_β is the short circuit, common emitter current gain cutoff frequency. Thus

$$\beta = \frac{G_d}{G_1}$$
$$\omega_\beta = \frac{G_1}{C_1} \tag{33}$$

Since for all reasonable transistors $\beta \gg 1$, in all cases of interest $G_d \gg G_1$. If the transistor is inverted, i.e., the collector forward biased (acting as an emitter) and the emitter reverse biased (acting as a collector) as shown in Fig. 10(b), $\rho_1 = -p_n =$ constant and no *a-c* current flows through G_1 or C_1.

The *a-c* emitter current

$$i_e = \rho_2 G_d \tag{34}$$

and the base current

$$i_b = \rho_2(G_2 + C_2 s). \tag{35}$$

Hence

$$\frac{i_e}{i_b} = \frac{G_d}{G_2 + C_2 s} = \frac{G_d/G_2}{1 + s\frac{C_2}{G_2}} = \frac{\beta_i}{1 + \frac{s}{\omega_{\beta i}}} \tag{36}$$

where β_i is the inverted current gain and $\omega_{\beta i}$ is the inverted cutoff frequency. Thus

$$\begin{aligned} B_i &= \frac{G_d}{G_2} \\ \omega_{\beta i} &= \frac{G_2}{C_2} \end{aligned} \tag{37}$$

Since β_i is often quite small, we may make no statement with regard to the relative magnitude of G_2 and G_d.

By the four simple measurements of β, β_i, ω_β and $\omega_{\beta i}$, we are able to determine all of the input elements in terms of one (preferably G_d). For many calculations, we need not proceed further. However, if we are interested in the voltage across a forward-biased junction, we need to determine $p_n\, G_d$. As in the case of the diode, we cannot determine either p_n or G_d separately by external measurements. Perhaps the best method of determining $p_n\, G_d$ is to measure the *d-c* emitter-base voltage v_{eb} and the *d-c* collector current I_c in the normal bias connection.

$$\begin{aligned} I_c &= (\rho_1 + \rho_n) G_d = p_n\, G_d\, e^{qv_{eb}/kT} \\ p_n\, G_d &= I_c\, e^{-qv_{eb}/kT} \end{aligned} \tag{38}$$

For germanium transistors, it is possible to obtain an approximate value of $p_n\, G_d$ from a measurement of i_{co}, the collector cutoff current when the emitter is open-circuited.

$$\begin{aligned} i_{co} &= p_n \left(G_2 + \frac{G_1 G_d}{G_1 + G_d} \right) = p_n\, G_d \left(\frac{1}{\beta_i} + \frac{1}{1+\beta} \right) \\ &\approx p_n\, G_d \left(\frac{1}{\beta_i} + \frac{1}{\beta} \right) \end{aligned} \tag{39}$$

hence

$$p_n\, G_d \approx \frac{i_{co} \beta \beta_i}{\beta + \beta_i} \tag{40}$$

However, i_{co} is normally composed of a certain amount of surface leakage current and junction depletion layer generation current. Hence, the value of $p_n\, G_d$ obtained in this way may be considerably in error. A measurement of v_{eb} and i_c at a bias current large compared to i_{co} is much to be preferred. In silicon units, the i_{co} is largely determined by carrier generation within the junction depletion region and therefore should never be used in the determination of $p_n\, G_d$.

Transistor Small Signal Performance. We have already used the common emitter current gain characteristic of the transistor in order to determine the values of the lumped model elements. It is of interest to investigate the other aspects of small signal performance as predicted by the lumped model. If the transistor is used in the common base connection and normally biased, the *a-c* collector current

$$i_c = \rho_1 G_d \tag{41}$$

and the *a-c* emitter current

$$i_e = \rho_1 (G_d + G_1 + C_1 s) \tag{42}$$

Therefore

$$\frac{i_c}{i_e} = \frac{G_d}{G_d + G_1 + C_1 s} = \frac{\frac{G_d}{G_d+G_1}}{1 + \frac{C_1}{G_d+G_1}} = \frac{\alpha}{1 + \frac{s}{\omega_\alpha}} \tag{43}$$

where

$$\begin{aligned} \alpha &= \frac{G_d}{G_d + G_1} = \frac{\beta}{1+\beta} \\ \omega_\alpha &= \frac{G_d + G_1}{C_1} = (1+\beta)\omega_\beta \end{aligned} \tag{44}$$

Thus, the lumped model gives the single time constant approximation for the current gain which is quite accurate for operation well below the alpha cutoff frequency, and is commonly used for high-frequency calculations. If $\beta \gg 1$, we may simplify the last expression as follows:

$$\omega_\alpha \approx \frac{G_d}{C_1} = \beta\omega_\beta \tag{45}$$

In the inverse common base connection, similar expressions apply.

$$\begin{aligned} i_e &= \rho_2 G_d \\ i_c &= \rho_2 (G_d + G_2 + C_2 s) \end{aligned}$$

$$\begin{aligned} \frac{i_e}{i_c} &= \frac{G_d}{G_d + G_2 + C_2 s} = \frac{\frac{G_d}{G_d+G_2}}{1 + \frac{C_2}{G_d+G_2} s} = \frac{\alpha_i}{1 + \frac{s}{\omega_{\alpha i}}} \\ \alpha_i &= \frac{G_d}{G_d + G_2} = \frac{\beta_i}{1+\beta_i} \\ \omega_{\alpha i} &= \frac{G_d + G_2}{C_1} = (1+\beta_i)\,\omega_{\beta i} \end{aligned} \tag{46}$$

Where β_i is not necessarily large compared to unity.

We may now ask what type of complete common emitter, small signal equivalent circuit results from the lumped model. For small signals, the emitter *a-c* minority carrier density is proportional to the *a-c* emitter-base voltage as shown in Eq. 18.

$$\rho_1 = \frac{q}{kT}\rho_o \, v_{be} = K_1 \, v_{be} \tag{47}$$

The *effective* *a-c* density at the collector is also proportional to the *a-c* collector-base voltage, due to collector depletion layer widening, or Early effect [2], [6].

$$\rho_2 = K - 2v_{cb} \qquad K_2 \ll K_1 \tag{48}$$

where all quantities of interest are now *a-c* components.

In the common emitter connection, expressions for the base and collector currents become

$$\begin{aligned} i_c &= -\rho_1 G_d + \rho_2(G_d + G_2 + C_2 s) \\ i_b &= \rho_1(G_1 + C_1 s) + \rho_2(G_2 + C_2 s) \end{aligned} \tag{49}$$

Since the emitter voltage is taken as zero, we may rewrite the currents in terms of collector and base voltages.

$$\begin{aligned} i_c &= v_{be}\left[-K_1 G_d + K_2(G_d + G_2 + C_2 s)\right] - K_2(G_d + G_2 + C_2 s)v_c \\ i_b &= -v_{be}\left[K_1(G_1 + C_1 s) - K_2(G_2 + C_2 s)\right] + K_2(G_2 + C_2 s)v_c \end{aligned} \tag{50}$$

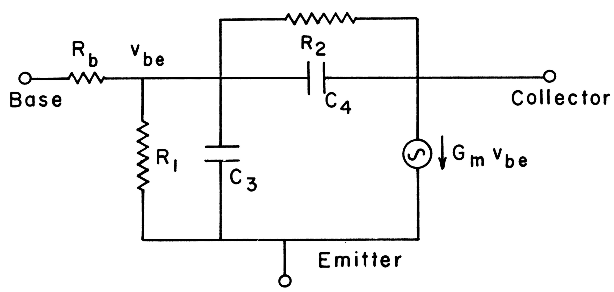


Figure 11: Transistor small signal equivalent circuit as developed from lumped model.

These equations correspond to the circuit shown in Fig. 11.

where

$$\begin{aligned} R_1 &= \frac{1}{K_1 G_1} \qquad & R_2 &= \frac{1}{K_2 G_2} \\ C_3 &= K_1 C_1 \qquad & C_4 &= K_2 C_2 \\ g_m &= K_1 G_d \end{aligned} \tag{51}$$

The extrinsic base resistance R_b must be added in series with the base terminal as shown. This circuit is similar to that proposed by Giacoletto [7] and widely used for high-frequency work. Thus, the lumped model reduces simply to a quite accurate representation in this special case, yet is much more general in that it is useful for all transistor operating conditions.

1.7 Transistor Switching Performance

It was recognized early in the development of the junction transistor that the symmetry of the device implied the unique ability to *saturate.* Since both the collector and emitter junctions are capable of emitting minority carriers into the base region when forward biased, and since the diffusion current across the base is due to the difference in minority carrier density between the two junctions, it is clear that when the transistor is saturated, *current may flow in either direction*, depending upon which density is the larger. In the saturated condition, the transistor closely resembles a closed switch, and dynamic resistances of a few tenths of an ohm are easily obtained with small units. If, on the other hand, both junctions are reverse biased, the transistor is cut off and only the very small junction reverse currents flow. Hence, the transistor resembles an open circuit, impedances of several megohms being common. When a transistor spends the majority of its time in one of two states, fully saturated or fully cut off, passing through the normally-biased region only to get from one state to the other, it is said to be operating as a *switch.* A distinction should be drawn at this point between true switching service and non-saturating service. A so-called non-saturating switch is one where the transistor may become either cut off or normally biased (usually with a rather low collector voltage), but not saturated. Such operation is more properly termed Class C pulse amplifier service. When the transistor is not caused to saturate, its operation may be analyzed with adequate accuracy by the use of small signal-equivalent circuits. However, when the transistor is caused to saturate, the analysis becomes quite complicated, and has traditionally been avoided in circuit work. Here, the lumped model comes into its own, since it transforms the difficult problem of transistor saturation and storage into one of simple R-C circuit analysis. The advantage of this approach for the circuit engineer is obvious.

Quasi-static Performance. Consider a transistor connected in the circuit shown in Fig. 12.

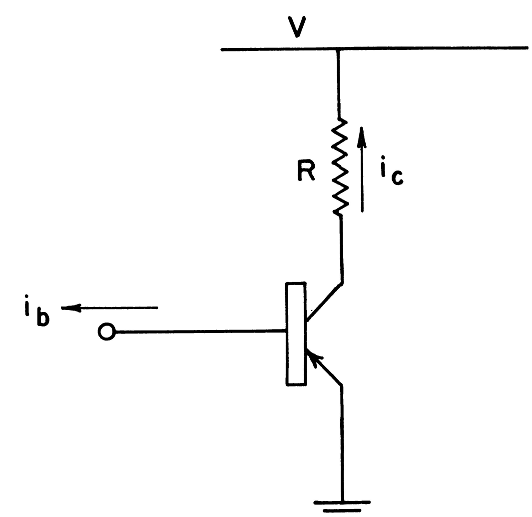


Figure 12: Elementary transistor switch.

For this analysis, we shall use the lumped model of Fig. 9. However, we may ignore C_1 and C_2, since we are only interested in slowly varying *d-c* quantities. The three regions of transistor operation will now be considered:

(a) Cut-Off. In the cut-off state, $\rho_1 = \rho_2 = -p_n$.

Therefore

$$i_c = p_n G_2 = \frac{i_{co}(1+\beta)}{1+\beta+\beta_i} \tag{52}$$

which becomes

$$i_c \approx \frac{i_{co}}{1 + \beta_i/\beta} \qquad \text{if} \qquad \beta \gg 1 \tag{53}$$

In the usual case, i_{co} is very small, and hence the collector voltage is approximately equal to the supply voltage

$$v_c \approx V \tag{54}$$

The base current

$$i_b = -p_n(G_1 + G_2) \approx -i_{co} \qquad \text{if} \qquad \beta \gg 1 \tag{55}$$

In this state, the only significant contribution to the power dissipation of the transistor is from the collector.

$$P \approx i_c V = \frac{i_{co} V}{1 + \beta_i/\beta} \tag{56}$$

(b) Active Region. In the active region

$$\rho_2 = -p_n \approx 0 \qquad \text{if} \qquad i_c \gg i_{co} \tag{57}$$

The collector excess density may be considered zero, provided the collector current is large compared with i_{co}. Thus

$$\begin{aligned} i_c &= \beta i_b \\ v_c &= V - \beta i_b R \end{aligned} \tag{58}$$

Since $\beta \gg 1$ and the collector voltage is larger than the base voltage (since the transistor is not yet saturated), the power dissipated due to base current is negligible compared with that due to the collector current. Thus

$$P \approx \beta V i_b - \beta^2 R {i_b}^2 \tag{59}$$

which reaches a maximum when $v_c \approx V/2$.

(c) Saturation. When the base current is increased to the point where $v_c = v_b$, the transistor saturates and the collector current becomes substantially independent of further increases in base current. Since both junctions are forward biased, the base and collector voltages will be neglected in comparison with V. Hence

$$\begin{aligned} i_c &\approx V/R \\ i_b &> V/\beta R \end{aligned} \tag{60}$$

We may now solve for the excess densities ρ_1 and ρ_2 in order to obtain the junction voltages v_{be} and v_{bc}. We may write the equations for base and collector current as

$$\begin{aligned} i_c &= (\rho_1 - \rho_2)G_d - \rho_2 G_2 \\ i_b &= \rho_1 G_1 + \rho_2 G_2 \end{aligned} \tag{61}$$

which may be solved for ρ_1 and ρ_2 writing β for G_d/G_1 and β_i for G_d/G_2

$$\rho_1 G_d \approx \frac{(1+\beta_i)i_b + i_c}{1 + \frac{\beta_i}{\beta} + \frac{1}{\beta}} \tag{62}$$

$$\rho_2 G_d \approx \frac{\beta_{ib} - i_c}{1 + \frac{\beta}{\beta_i} + \frac{1}{\beta_i}} \tag{63}$$

The combination ρG_d is a convenient quantity with which to deal since it has the dimensions of a current. When $\beta i_b = i_c$, the transistor has just become saturated and

$$i_c = \rho_1 G_d = V/R \tag{64}$$

At higher base currents, the collector current remains essentially constant, but $\rho_2 G_d$ increases. Hence, $\rho_2 G_d$ is a significant measure of the amount by which the transistor has been driven into saturation. We may now determine the junction voltage since

$$\begin{aligned} \rho_1 G_d &= p_n G_d (e^{qv_{be}/kT} - 1) \\ \rho_2 G_d &= p_n G_d (e^{qv_{bc}/kT} - 1) \end{aligned} \tag{65}$$

Therefore

$$\begin{aligned} v_{be} &= \frac{kT}{q} \mathrm{Ln} \left(\frac{\rho_1 G_d}{p_n\, G_d} + 1 \right) \\ v_{bc} &= \frac{kT}{q} \mathrm{Ln} \left(\frac{\rho_2 G_d}{p_n\, G_d} + 1 \right) \end{aligned} \tag{66}$$

where $p_n\, G_d$ may be determined from Eq. 38 or Eq. 40. Although the i_{co} expression is not as accurate as a measurement of junction voltage and current, it is often convenient for germanium units.

$$\begin{aligned} v_{be} &\approx \frac{kT}{q} \mathrm{Ln} \left[\frac{i_c + (1 + \beta_i) i_b}{\beta_i i_{co}} + 1 \right] \\ v_{bc} &\approx \frac{kT}{q} \mathrm{Ln} \left[\frac{(1 + \beta) i_b - i_c}{\beta i_{co}} + 1 \right] \end{aligned} \tag{67}$$

The collector saturation voltage is just the difference between the two junction voltages. In most cases, the drive current is sufficiently large that ρ_1 and ρ_2 are both much greater than p_n. Thus from Eq. 67

$$v_c \approx \frac{kT}{q} \mathrm{Ln} \frac{\rho_1 G_d}{\rho_2 G_d} \tag{68}$$

which from Eq. 62 and Eq. 63 becomes

$$v_c = \frac{kT}{q} \mathrm{Ln} \left(\frac{1 + \frac{1 + i_c/i_b}{\beta i}}{1 - i_c/\beta i_b} \right) \tag{69}$$

In the inverted connection, the normal and inverse quantities merely exchange places and the junction voltages become

$$v_{bc} = \frac{kT}{q} \mathrm{Ln} \left[\frac{i_e + (1 + \beta) i_b}{\beta i_{co}} + 1 \right] \tag{70}$$

$$v_{be} = \frac{kT}{q} \mathrm{Ln} \left[\frac{(1+\beta_i)i_b - i_e}{\beta_i i_{co}} + 1 \right] \tag{71}$$

Also, the emitter saturation voltage becomes

$$v_e = \frac{kT}{q} \mathrm{Ln} \left(\frac{1 + \frac{1+i_e/i_b}{\beta}}{1 - i_e/\beta_i i_b} \right) \tag{72}$$

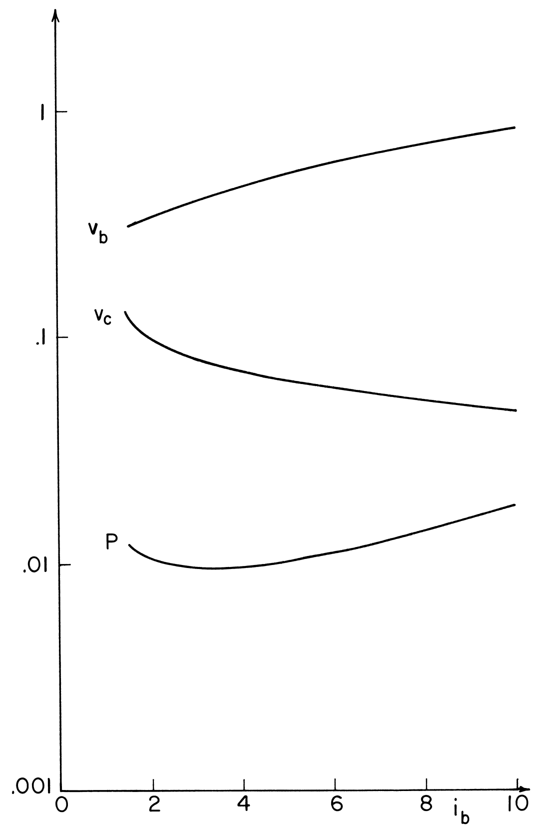


Figure 13: Voltage and power variation for normally connected transistor.

Plots of the base input voltage, saturation voltage, and power dissipation for a typical switching transistor operating in the normal and inverse connections are shown in Fig. 13 and Fig. 14. It is of interest to note the minimum in power dissipation which occurs at some base drive current. Clearly, this drive current represents an optimum operating point for the particular transistor and collector current involved. The serious nature of large overdrive currents is quite apparent.

Dynamic Resistance. The resistance of a saturated transistor to small *a-c* signals is often of interest. This dynamic resistance is given by

$$R_s = \frac{\delta v_c}{\delta i_c} \tag{73}$$

for the normal connection, and may be obtained from Eq. 69.

$$R_s = \frac{kT}{q} \left[\frac{1}{(1+\beta)i_b - i_c} + \frac{1}{(1+\beta_i)i_b + i_c} \right] \tag{74}$$

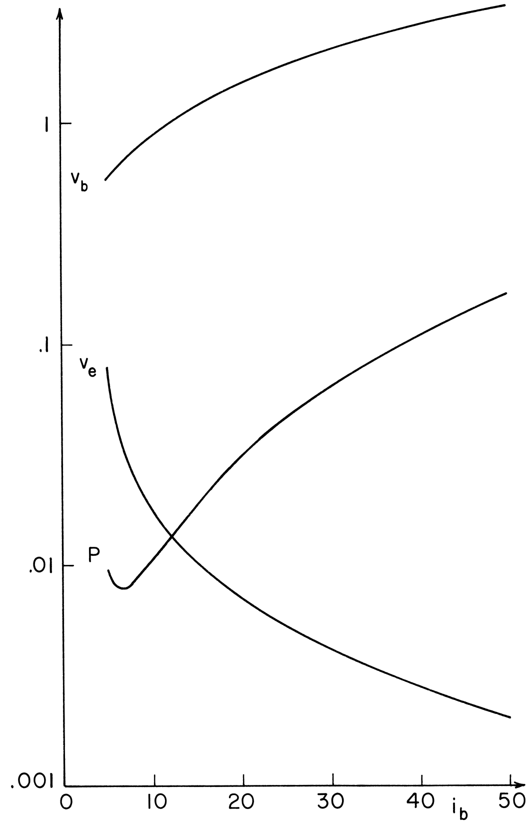


Figure 14: Voltage and power variation for inversely-connected transistor.

A similar expression is obtained for the inverted connection, only the roles of the collector and emitter are interchanged, and β is interchanged with β_i. For sufficiently large drive currents

$$R_s \approx \frac{kT}{q}\left[\frac{\beta+\beta_i}{\beta(1+\beta_i)i_b}\right] \qquad \beta \gg 1 \tag{75}$$

which shows the saturation resistance inversely dependent on β, β_i and the drive current. The importance of *both* β and β_i is dramatically illustrated by this formula.

All transistors have certain ohmic resistances associated with their collector and base circuits due to the semiconductor material between the active region of the device and its contact to the outside world. For the first approximation, we may assume these resistances constant, and if the voltage drops across them are not negligible, we must add them to the appropriate voltages already calculated. For switching transistors made by the alloying process, and many others, the collector series resistance is negligible, but the base resistance R_b should always be taken into account. The total base voltage thus becomes

$$v_b = v_{be} + i_b\, R_b \tag{76}$$

Since we are considering base currents up to very high values, we may no longer neglect the power dissipation in the base circuit.

$$P = v_c i_c + v_{be} i_b + {i_b}^2\, R_b \tag{77}$$

As we have noted, a number of simplifying assumptions have been made which under many conditions may be very questionable. However, for practical circuit design, one often uses models which are greatly oversimplified, not because of their extreme accuracy, but because they provide approximate answers and still allow a qualitative understanding of the problem. For example, the use of small signal equivalent circuits without regard to the magnitude of the signal level, is an accepted engineering procedure. The non-linearities are taken into account qualitatively after the main circuit behavior has been determined from the linear analysis. The lumped model serves in the same capacity for switching problems as a small signal equivalent circuit does for problems where the transistor is normally biased. It provides a straightforward method of obtaining results with reasonable accuracy in the majority of cases, and hence may claim a certain engineering importance. The physical insight gained by the lumped model analysis is often much more valuable than a slight improvement in accuracy, since it enables the analyst to make qualitative statements concerning changes in circuit parameters, a very important step in the design procedure.

Transient Response [8]. Again, we shall consider a junction transistor connected as shown in Fig. 12.

(a) Turn On. In the cut-off condition $i_b = -i_{co}$, as before. Now, let us apply a positive current step of magnitude I_1 (large compared to i_{co}) and calculate the collector current response. As long as the transistor remains normally biased, the collector current may be computed from small signal formulae

$$i_c(s) = \frac{\beta i_b(s)}{1 + \frac{s}{\omega\beta}} = \frac{\beta\omega_\beta I_1}{s(s + \omega_\beta)}$$
$$i_c = \beta I_1(1 - e^{-w_\beta t}) \tag{78}$$

However, the collector remains reverse biased only so long as

$$i_c < V/R \tag{79}$$

Thus, the collector current rises toward the asymptote βI_1 with time constant $1/\omega\beta$. If $\beta I_1 > V/R$, the transistor will saturate when the collector current reaches V/R. The "rise time" required is thus

$$t_r = -\frac{1}{\omega_\beta}\ln\left(1 - \frac{V}{\beta I_1\, R}\right) \tag{80}$$

Usually, the transistor is driven quite hard in order to minimize the rise time. Under these conditions, $\beta I_1 \gg V/R$, and we may expand the log, retaining only the first term

$$t_r \approx \frac{V}{\beta\omega_\beta\, RI_1} \tag{81}$$

Since, normally, $\beta \gg 1$, $\omega_\alpha \approx \beta\omega_\beta$. The rise time may be written

$$t_r \approx \frac{V}{\omega_\alpha RI_1} \tag{82}$$

It is thus clear that the alpha cut-off frequency is the determining factor in turn-on time, and not β or ω_β alone. The conditions during turn-on are illustrated in Fig. 15.

(b) Storage. After the transistor has reached the steady state with $i_b = I_2$, let us suddenly reverse the base current to $i_b = -I_3$. As in the case of the diode, the collector junction remains forward biased, since ρ_2 cannot change instantaneously. Hence, the collector current remains $I_c = V/R$. After a "storage time,"

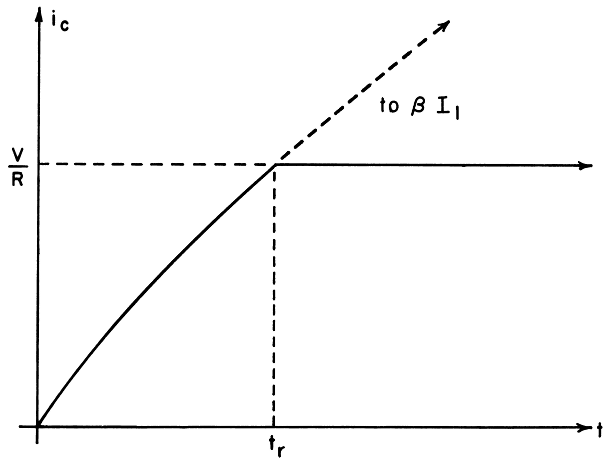


Figure 15: Rise transient as predicted by transistor lumped model.

t_s, ρ_2 has reached zero, and the transistor becomes normally biased [10]. In order to determine ρ_1 and ρ_2 during the storage period, we may determine their initial values from the steady-state conditions given by Eq. 63, assuming $\beta \gg 1$

$$\rho_2(0) = \frac{\beta I_2 - I_c}{\frac{\beta}{\beta_i} + 1} \qquad\qquad I_c = V/R \tag{83}$$

The transient densities may be found by superposing the steady-state solution above (for $I_c = V/R$, $i_b = I_2$) upon the solution for a negative base current step of magnitude $I_2 + I_3$, and constant collector current $i_c = 0$.

The result of this calculation, assuming $\beta \gg 1$, is

$$\rho_2 G_d \approx (I_2 + I_3)\frac{\beta\beta_i}{\beta + \beta_i}\, e^{-bt} - \frac{\beta_i I_c + \beta\beta_i I_3}{\beta + \beta_i} \tag{84}$$

where $I_c = V/R$ and

$$b \approx \frac{\omega_\alpha \omega_{\beta i} + \omega_\beta \omega_{\alpha i}}{\omega_\alpha + \omega_{\alpha i}} \tag{85}$$

The "storage time" t_s, ends when ρ_2 reaches zero.

Hence

$$e^{-bt_s} = \frac{\beta_i I_c + \beta\beta_i I_3}{\beta\beta_i(I_2 + I_3)} \tag{86}$$

or

$$t_s = \frac{1}{b} \ln \frac{I_2 + I_3}{I_3 + \frac{I_c}{\beta}} \tag{87}$$

from which it is clear that the storage time may be reduced by using large turn-off currents. I_2 must be greater than I_c/β for the transistor to be saturated, but the overdrive may be reduced to decrease the storage time. *The controlling time constant for the storage period is* b, hence for small storage times, *both* ω_β and $\omega_{\beta i}$ should be large.

If these frequencies are nearly equal

$$b \approx \omega_\beta \approx \omega_{\beta i} \tag{88}$$

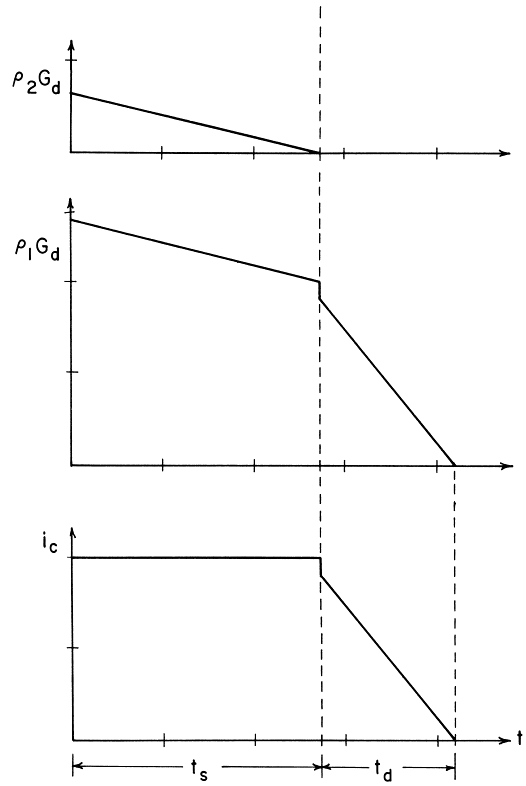


Figure 16: Storage and decay transients as predicted by transistor lumped model.

Plots of ρ_1, ρ_2 and i_c during the storage period are shown in Fig. 16.

(c) Turn Off. When ρ_2 approaches zero, the collector current is made up of two components:

$$I_c = \rho_1 G_d + C_2 \frac{d\rho_2}{dt} \tag{89}$$

However, as the collector junction becomes reverse biased, $d\rho_2/dt$ abruptly becomes zero since ρ_2 cannot become less than $-p_n$, and the lumped model predicts a slight discontinuity in collector current. In reality, the actual current changes smoothly, and this is another case where the lumped nature of the model fails to give the completely correct physical picture. However, for purposes of calculating the "decay time," or time required for i_c to reach zero, the lumped model expressions, including the discontinuity, will be more accurate than the corresponding expressions, assuming no discontinuity. The reason is that the true collector current very quickly approaches the predicted value as an asymptote, and by the time i_c approaches zero, the lumped model expression is quite accurate. Let us, therefore, calculate the magnitude of the discontinuity.

$$\Delta i = C_2 \frac{d\rho_2}{dt}\bigg|_{t=t_s} = -\frac{C_2}{G_d} b \, \frac{(\beta_i I_c + \beta\beta_i I_3)}{\beta + \beta_i} \tag{90}$$

Substituting for b and assuming $\beta \gg 1$

$$\Delta i = -(I_c + \beta I_3)\left(\frac{\omega_\beta}{\omega_\alpha + \omega_{\alpha i}}\right) \tag{91}$$

Thus, under normal circumstances, the relative magnitude of the discontinuity is quite small. However, for very large overdrive currents, it may become important. Under these conditions

$$\Delta i \approx -\beta I_3 \frac{\omega_\beta}{\omega_\alpha + \omega_{\alpha i}} \tag{92}$$

After ρ_2 reaches zero, the transistor is again normally biased, and we may use the small signal approach, as with the turn-on period. The collector current approaches the asymptote $-\beta I_3$ with a time constant $1/\omega_\beta$. During this period, $i_c = \rho_1 G_d$

$$i_c = (I_c - \Delta i)e^{-\omega_\beta t} - \beta I_3(1 - e^{-\omega_\beta t}) \tag{93}$$

The "decay time" t_d is the time required for i_c to reach zero

$$t_d = \frac{1}{\omega_\beta} \ln\left(1 + \frac{I_c + \Delta i}{\beta I_3}\right) \tag{94}$$

For large turn-off current $I_3 \gg I_c/\beta$, and we may approximate the logarithm. The decay time therefore becomes

$$t_d \approx \frac{1}{\omega_\alpha}\left(\frac{I_c}{I_3} - 1 + \frac{1}{\omega_{\alpha i}/\omega_\alpha}\right) \qquad \text{if} \qquad \beta \gg 1 \tag{95}$$

Thus, as with the rise time, the *decay time is determined by the magnitude of the drive current and the α cut-off frequency*, and not by β or ω_β alone.

After the turn-off period, the transistor is cut off and the collector current resumes its small steady-state value as given by Eq. 53. Comparing the expression for rise time as given in Eq. 82, we see that the decay time is always less than the rise time for a given base-drive current. This is true because recombination is *helpful* during the decay period, but *harmful* during the rise period.

References

[1] Ebers, J.J. and J.L. Moll, "Large Signal Behavior of Junction Transistors," Proc. IRE, **42**, p. 1761, (December, 1954).

[2] Middlebrook, R.W., "An Introduction to Junction Transistor Theory," John Wiley and Sons, Inc., New York, (1957).

[3] Linvill, J.G., "Lumped Models of Transistors and Diodes," Proc. IRE, **46**, p. 1141, (June, 1958).

[4] Moll, J.L., "The Evolution of the Theory for the Voltage-Current Characteristics of *P-N* Junctions," Proc. IRE, **46**, p. 1076, (June, 1958).

[5] Kingston, R.H., "Switching Time in Junction Diodes and Junction Transistors," Proc. IRE, **42**, p. 829, (May, 1954).

[6] Early, J.M., "Effects of Space Charge Layer Widening in Junction Transistors," Proc. IRE, **40**, p. 1401, (November, 1952).

[7] Giacoletto, L.J., "Study of *P-N-P* Alloy Junction Transistors from D-C through Medium Frequencies," RCA Review, **15**, p. 506, (December 1954).

[8] Moll, J.L., "Large Signal Transient Response of Junction Transistors," Proc. IRE, **42**, p. 1773, (December, 1954).

[9] Hurley, R.B., "Junction Transistor Electronics," John Wiley and Sons, Inc., New York, Chapter 19, (1958).

[10] DeWitt, D. and A.L. Rossoff, "Transistor Electronics," McGraw-Hill Book Co., Inc., New York, Chapter 11, (1957).