1. WE WANT $I_E = I_c$ so that the fundamental basis of transistor action, that of a current controlled by a small voltage flowing across a large resistor to generate a large voltage is maintained.

This is not true because,

a. Emitter Injection efficiency, $\alpha$ . $I_E$ is composed to two current components, $I_{Ep}$ and $I_{En}$ . $I_{Ep}$ is desirable because it is the component that injects minority carriers into the B-C junction and emerges as $I_c$ . $I_{En}$ is the electron current injected into the emitter from the base. This is a parasitic current as it does not contribute to collector current, $I_c$ . Therefore we define a figure of merit (which we will calculate shortly) called EMITTER INJECTION EFFICIENCY, $\alpha$ , where

$$\alpha = \frac{I_{Ep}}{I_{Ep} + I_{En}}$$

The maximum (and desirable value of $\alpha$ is 1 and is achieved when $I_{En} \rightarrow 0$.

b. BASE TRANSPORT FACTOR, $\alpha_T = \frac{I_{Ep}}{I_{Ep}}$ . Of the hole flux emitted from the emitter, a certain fraction is lost due to recombination in the n-type base. However, if the width of the neutral base, $W_B$ is made much smaller than the average distance a minority carrier travels before recombining, $L_p$ , then the probability of recombination is reduced.
So what does the charge profile in the base region look like? USE THE CONTINUITY EQUATION. In the absence of electric field this reduces to the diffusion equation

\[ D_p \frac{d^2 p_n}{dx^2} = \text{Net recombination} \equiv 0 \text{ if } W_B \ll L_p \]

Use a coordinate system shown in the figure above where the neutral base commences at \( x = 0 \) and ends \( x = W_B \). The solution of

\[ D_p \frac{d^2 p_n}{dx^2} = 0 \text{ is the same as } D_p \frac{d^2 \Delta p_n(x)}{dx^2} = 0 \text{ and is } \Delta p_n(0) \left[ 1 - \frac{x}{W_B} \right]. \]

This is a linear decrease in the charge profile with boundary conditions that \( \Delta p_n @ x = 0 \) is \( \Delta p_n(0) \). Therefore the charge profile in the bipolar transistor can be pictured as

\[ \Delta n_p(x') = \Delta n_p(0)e^{-\frac{x'}{L_n}} \]

\[ p_n(x) = \Delta p_n(0) \left[ 1 - \frac{x}{W_B} \right] \]

The salient features are

(i) An exponential drop-off in the injected minority profile as the emitter thickness is much larger than a diffusion length for electrons

\[ \Delta n_p(x') = \Delta n_p(0)e^{-\frac{x'}{L_n}} \text{ [like a conventional diode]} \]

(ii) A linear decrease in the minority (hole concentration in the base

\[ \Delta p_n(x) = \Delta p_n(0) \left[ 1 - \frac{x}{W_B} \right] \text{ because we neglected recombination.} \]

(iii) A traditional decrease in the minority carrier concentration \( n_p(x) \) in the collector as determined by reverse biased conditions \( n_p(x^c) = n_{p0}\left(1 - e^{-\frac{x^c}{L_n}}\right) \)

NOW WE CAN CALCULATE ALL THE CURRENTS BECAUSE WE KNOW THE CHARGE PROFILES.

Recall that we want \( I_c = I_E \) which requires \( \alpha = 1 \) and \( \alpha = 1 \). In general the current gain \( \alpha \) is defined as \( \alpha = \alpha_r \cdot \alpha \) and \( I_c = \alpha I_E \) WE WANT \( \alpha = 1 \)
CALCULATING $\alpha$ \quad (A = 1 cm$^2$)

\[ \alpha = \frac{I_{Ep}}{I_{Ep} + I_{EN}} = \frac{1}{1 + \frac{I_{EN}}{I_{Ep}}} \]

\[ I_{Ep} = qD_p \left. \frac{dn_p}{dx} \right|_{x=0} \quad I_{EN} = qD_n \left. \frac{dn_p}{dx} \right|_{x=0} \]

\[ I_{Ep} = qD_p \frac{\Delta n_p(0)}{W_B} \quad I_{En} = qD_n \frac{\Delta n_p(0)}{L_n} \]

OR \quad \[ I_{Ep} = qD_p \left( \frac{p_n(0)}{W_B} \left( e^{\frac{qV_n}{kT}} - 1 \right) \right) \quad I_{En} = qD_n \left( n_p(0) \left( e^{\frac{qV_n}{kT}} - 1 \right) \right) \]

\[ \therefore \quad \frac{I_{En}}{I_{Ep}} = \frac{D_n}{D_p} \cdot \frac{n_p(0)}{p_n(0)} \cdot \frac{W_B}{L_n} \]

Using the Law of Mass Action

\[ n_p(0) = \frac{n_i^2}{N_{AE}} \quad & \quad p_n(0) = \frac{n_i^2}{N_{DB}} \]

where \( N_{AE} \) = Acceptor concentration in the p-type emitter

& \( N_{DB} \) = Donor concentration in the base

\[ \therefore \quad \frac{I_{En}}{I_{Ep}} = \frac{D_n}{D_p} \cdot \frac{N_{DB}}{N_{AE}} \cdot \frac{W_B}{L_n} \]

For \( \alpha \to 1 \) we need \( \frac{I_{En}}{I_{Ep}} \to 0 \)

\[ \therefore \quad \text{We Need} \]

<table>
<thead>
<tr>
<th>$\frac{W_B}{L_n}$</th>
<th>$&lt;&lt; 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{N_{AE}}{N_{DB}}$</td>
<td>$&gt;&gt; 1$</td>
</tr>
</tbody>
</table>

Calculating BASE TRANSPORT FACTOR, $\alpha_T$.

\[ \alpha_T = \frac{I_{Ep}}{I_{Ep}} = \frac{\text{collected hole current}}{\text{injected hole current}} \]

Also \( \alpha_T = \frac{I_{Ep} - I_{Br}}{I_{Ep}} \)

where \( I_{Br} \) is the hole current lost due to recombination in the base.
We assumed for calculating the hole profile in the base the recombination was zero. This gave us a linear profile. In actuality, the finite (through very small) recombination that does occur (as $W_B/L_p < 1$ but not 0) causes the hole profile to sag as shown below because holes are lost to recombination.

**Bottom line:** the loss of holes perturbs the profile a bit but it is essentially linear therefore validating our assumption.

Or Neglecting recombination is fine to calculate the collector current as the slope of the profile (which determines $I_{cp}$) is not perturbed substantially.

i.e. $I_{cp} = I_{Ep} = qD_p \frac{\Delta p_n(0)}{W_B}$

But to calculate $I_{Br}$ we have to account for recombination. This is done elegantly using charge control analysis.

$I_{Br}$, the recombination current arises from the re-supply of electrons in the base lost to recombination with the holes. The rate at which holes recombine is given by $\frac{Q_p}{\tau_p}$.

$Q_p$ can be calculated as $Q_p = \frac{1}{2} \cdot \Delta p_n(0) \cdot W_B$
\[ I_{Br} = \frac{Q_p}{\tau_p} = \frac{1}{2} p_n \left( \frac{q V_T}{\tau_p} - 1 \right) \frac{W_B}{\tau_p} \]

We now define current gain, \( \beta \), to be \( \beta = \frac{I_c}{I_B} = \frac{I_{cp}}{I_{En} + I_{Br}} \) neglecting \( I_{En} \) (by properly designing \( \delta \rightarrow 1 \)) gives us \( \beta = \frac{I_{cp}}{I_{Br}} \). But \( I_{cp} \equiv I_{Ep} = qD_p \frac{\Delta p_n(0)}{W_B} \)

\[ \beta = \frac{I_{cp}}{I_{Ep}} = \left( qD_p \frac{\Delta p_n(0)}{W_B} \right) \frac{1}{2} \frac{\Delta p_n(0)}{\tau_p} \frac{W_B}{\tau_p} \]

\[ \beta = 2 \frac{D_p \tau_p}{W_B^2} \]

\[ \beta = 2 \frac{I_p^2}{W_B^2} \] as \( D_p \tau_p = I_p \)

for \( \delta \rightarrow 1 \)

The relationship between \( \beta \) and \( \alpha \)

\[ \beta = \text{COMMON-EMITTER CURRENT GAIN} \]

\[ \alpha = \text{COMMON-BASE CURRENT GAIN} \]

\[ \begin{align*}
E & \quad \text{Input voltage between Emitter and Base} \\
C & \quad \text{Output voltage measured between Collector and Base} \\
V_{in} & \quad \text{Input voltage between Emitter and Base} \\
V_{out} & \quad \text{Output voltage measured between Collector and Base} \\
\end{align*} \]

\[ \text{BASE is the common terminal} \]

\[ I_{in} = I_E; \quad I_{out} = I_c \]

\[ I_c = \alpha I_E \]

Note:
\[ \alpha = \alpha_t \cdot \delta \]

If \( \delta \rightarrow 1 \)
then \( \alpha = \alpha_t \)

Common Emitter Configuration
Here \( V_{in} \) is applied between the base and emitter and the output voltage measured between collector and emitter. The emitter is the common terminal. The input current, \( I_{in} = I_B \) and the output current, \( I_{out} = I_c \)

\[ \therefore \frac{I_{out}}{I_{in}} = \frac{I_c}{I_B} = \beta \]
In general

\[ I_E - I_B = I_c \] (Kirchoff's Law)

using \( I_E = \frac{I_c}{\alpha} \) & \( I_B = \frac{I_c}{\beta} \)

\[ \frac{I_c}{\alpha} = \frac{I_c}{\beta} \]

\[ \frac{1}{\alpha} = \frac{1}{\beta} = 1 \]

OR \( \beta = \frac{\alpha}{1-\alpha} \) Typically: \( \beta > 10, \alpha > 0.9 \)

Therefore if \( \delta = 1 \) then since \( \alpha = \alpha_r \) we get from \( \beta = \frac{\alpha}{1-\alpha}, \ \alpha = \frac{\beta}{\beta + 1} \equiv \alpha_r \)

\[ \therefore \alpha_r = \frac{2 \frac{L_p^2}{W_p^2}}{1 + \frac{2L_p^2}{W_p^2}} = \frac{2L_p^2}{2L_p^2 + W_p^2} \]

as \( W_B \to 0, \alpha_r \to 1 \) as expected because recombination \( \to 0 \)

**SUMMARY DESIGNING GOOD BIPOLAR TRANSISTORS**

\( \delta \to 1 \Rightarrow \frac{N_{\text{mix}}}{N_{\text{BB}}} \gg 1 \)

\[ \frac{W_B}{L_p} \ll 1 \]

\( \alpha \to 1 \Rightarrow \frac{W_B}{L_p} \ll 1 \)

**DECREASE THE BASE THICKNESS, INCREASE THE BASE MINORITY CARRIER LIFETIME, INCREASE EMITTER DOPING.**
Addendum to the $f_{max}$ derivation

There was a question in class about the output resistance. There are two sources of output resistance. At D.C. it is $\frac{V_A}{I_c}$. At high frequencies the following derivation is valid which gives a real output impedance in addition to $\frac{V_A}{I_c}$.

$Z_{in}$ can be any value

![Circuit Diagram]

We assume that the frequency is high enough so that $\frac{1}{j\omega C_{in}} \ll r_b$ and the current generator

$$
\left( \frac{\beta_o}{1 + \frac{j\omega \beta_o}{w_T}} \right)^b \rightarrow \frac{j\omega + i_b}{w}
$$

$i_b$ is the current flowing through $C_a$ which is $[C_T + C_{be}]$, the total input capacitance.

Since $r_b \gg \frac{1}{j\omega C_{in}}$ the input termination is irrelevant (and the current $r_b$ is neglected).

∴ On applying a test voltage source $V_o$

The output impedance is $Z_o = \frac{V_o}{I_o}$.

$$I_o = I_{gen} + i_b$$

Assume $C_{bc} \ll C_{in}$

Then $i_b = j\omega C_{bc} (V_o - V_{be})$

$$= j\omega C_{bc} V_o$$
\[ \therefore I_o = jw C_{bc} V_o + I_{gen} \]

\[ = jw C_{bc} V_o + \left( -\frac{jw_T}{w} \right) I_b \]

\[ = jw C_{bc} V_o + \frac{w_T}{w} \cdot C_{bc} V_o \cdot \varepsilon \]

\[ \therefore \text{ The output impedance is} \]

\[
\begin{array}{c}
\frac{1}{w_T C_{bc}} \\
\hline
C
\end{array}
\]

The \textit{D.C.} conductance of \( \frac{I_c}{V_A} \) is in parallel, but is a small conductance.

\textbf{HETEROJUNCTIONS:}
Abrupt \textit{p–n} junctions:

\[ V_{D1} \quad \varepsilon_1 \quad \varepsilon_2 \quad V_{D2} \]

\[ \varepsilon_1 E_1 = \varepsilon_2 E_2 \]
\[
\begin{align*}
V_{D1} + V_{D2} &= V_{bi} \\
\frac{1}{2} \varepsilon_1 x_n &= V_{D1} \\
\frac{1}{2} \varepsilon_2 x_p &= V_{D2} \\
\Rightarrow V_{bi} &= \frac{1}{2} \varepsilon_1 x_n + \frac{1}{2} \varepsilon_2 x_p \\
&= \frac{1}{2} \varepsilon_1 x_n + \frac{1}{2} \varepsilon_2 x_p
\end{align*}
\]

Charge neutrality \(\Rightarrow qN_D x_n = qN_A x_p \)

\[
\Rightarrow \frac{N_D}{N_A} = \frac{x_p}{x_n}
\]

Also \(\frac{qN_D}{\varepsilon_1} x_n \) \& \(\frac{qN_A}{\varepsilon_2} x_p\)

\[
V_{bi} = \frac{q}{2} \left\{ \frac{\varepsilon_2 N_D x_n^2 + \varepsilon_1 N_A x_p^2}{\varepsilon_1 \varepsilon_2} \right\}
\]

\[
N_D x_n = N_A x_p
\]

\[
\therefore x_n = \frac{N_A x_p}{N_D}
\]

\[
V_{bi} = \frac{q}{2} \left\{ \frac{\varepsilon_2 N_D \left( \frac{N_A}{N_D} x_p \right)^2 + \varepsilon_1 N_A x_p^2}{\varepsilon_1 \varepsilon_2} \right\}
\]

\[
= \frac{q}{2} \left\{ \frac{\varepsilon_2 \frac{N_A^2}{N_D} x_p^2 + \varepsilon_1 N_A x_p^2}{\varepsilon_1 \varepsilon_2} \right\}
\]

\[
= \frac{q}{2} \frac{N_A x_p^2 \left( \frac{\varepsilon_2}{N_D} + \frac{\varepsilon_1}{N_A} \right)}{\varepsilon_1 \varepsilon_2}
\]

\[
\therefore x_p = \frac{1}{N_A} \sqrt{\frac{2}{q} \frac{\varepsilon_1 \varepsilon_2 N_A N_D}{\varepsilon_2 N_A + \varepsilon_1 N_D} V_{bi}}
\]

\[
x_n = \frac{1}{N_D} \sqrt{\frac{2}{q} \frac{\varepsilon_1 \varepsilon_2 N_A N_D}{\varepsilon_2 N_A + \varepsilon_1 N_D} V_{bi}}
\]

**Homojunction case**: \(\varepsilon_1 = \varepsilon_2\)
\[ x_p = \frac{1}{N_A} \sqrt{\frac{2\varepsilon}{q} \left( \frac{N_A N_D}{N_D + N_A} \right) V_{bi}} \]

\[ = \sqrt{\frac{2\varepsilon}{q} \frac{N_D}{N_A (N_D + N_A)} V_{bi}} \]

\[ V_{bi} = \Phi_2 - \Phi_1 \]

\[ \Phi_2 = \chi_2 + \left( \frac{Eg_2}{2} + \Phi_p \right) \]

\[ \Phi_1 = \chi_1 + \left( \frac{Eg_1}{2} - \Phi_n \right) \]

\[ \therefore V_{bi} = \chi_2 - \chi_1 + \left( \Phi_p - \Phi_n \right) + \frac{1}{2} (Eg_2 - Eg_1) \]

\[ = \Delta \chi + \left( \Phi_p + \Phi_n \right) - \frac{1}{2} \Delta E_g \]

For homojunction \[ \Delta \chi = \Delta E_g = 0 \]

\[ V_{bi} = \Phi_p - \Phi_n \text{ as expected} \]

Relative voltages: always true

\[ V_{D_1} = \frac{1}{2} \epsilon_1 x_n \]

\[ V_{D_2} = \frac{1}{2} \epsilon_2 x_p \]

\[ \therefore \frac{V_{D_1}}{V_{D_2}} = \frac{\epsilon_2 N_D x^2_n}{\epsilon_1 N_A x^2_p} \]

\[ = \frac{\epsilon_2}{\epsilon_1} \frac{N_D x_n}{x_p} \cdot \frac{x_n}{x_p} \]

\[ = \frac{\epsilon_2}{\epsilon_1} \frac{x_n}{x_p} \]

\[ = \frac{\epsilon_2}{\epsilon_1} \frac{N_A}{N_D} \]

When you apply bias \[ V_a \]
\[ \frac{V_{d1} - V_1}{V_{d1} - V_2} = \frac{\varepsilon_2}{\varepsilon_1} \frac{N_A}{N_D} \quad \& \quad V_1 + V_2 = V_a \]

Solving the equations simultaneously gives us \(V_1\) & \(V_2\).

\(V_{d1} - V_1\) is now the barrier to electron flow. If one considers thermic emission then

\[ I \propto \exp \frac{qV_1}{kT} \]
ON THE THEORY OF DEBYE AVERAGING IN THE C-V PROFILING OF SEMICONDUCTORS

HERBERT KROEMER and WU-YI CHIEN
Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106, U.S.A.

(Received 26 August 1980, in revised form 7 November 1980)

Abstract- We discuss the nature of the Debye averaging that takes place in the C–V profiling of highly non-uniform electron distributions in semiconductors. It is shown that the averaging process preserves the moment of the electron distribution. This property can be used to extract certain quantitative details about the electron distribution that appear to have been obliterated by the averaging, without the need to reconstruct the entire true electron distribution. As an example, the extraction of the steepness of a diffusion gradient is discussed in the limit of a weak electron concentration gradient, the width of the Debye averaging is shown to be such that the RMS averaging distance is \( \sqrt{(2)L_D} \), where \( L_D \) is the Debye length.

I. Introduction
It has long been recognized [1,2] that C–V profiling is, conceptually, a measurement of the free electron concentration \( n_d(x) \) rather than of the doping concentration of semiconductors. Whenever space charges are present, the two are not the same. And space charges will in general be present whenever the measured electron concentration is not uniform.* Kennedy and O’Brein (3) (KOB) have described a correction that should permit one to obtain the exact doping concentration \( n_d(x) \) from the measured electron concentration \( n(x) \) if the latter were indeed correctly given by C–V profiling. However, it is also well known (4) that C–V profiling, even if viewed as a measurement of the free electron concentration, is itself inexact when \( n(x) \) is non-uniform. More specifically when the true electron concentration \( n(x) \) varies by an appreciable fraction, over a distance less than a Debye length.

\[ L_D = (\epsilon kT / \pi q^2 N)^{1/2}, \]  

(1)

When the measured apparent electron concentration \( \hat{n}(x) \) is an average of the true \( n(x) \), taken over a distance of the order. By apparent electron concentration, we refer back to the concentration that is obtained formally from the “canonical” interpretation formula of C–V profiling.

* Present address, Fairchild Camera and Instrument Corp., 101 Bernal Road, San Jose, CA 95119, U.S.A.

-More exactly, whenever \( \log n(x) \) depends on position in any other than linearly. This includes the ends of any finite linear range.
\[ n(w) = \left[ \frac{q \varepsilon}{2} \frac{d}{dV} \left( \frac{1}{C^2} \right) \right]^{-1} . \]  \eqref{2a}

where
\[ w = \varepsilon / C . \]  \eqref{2b}

Here \( \varepsilon \) is the permittivity of the semiconductor, \( C \) the differential capacitance per unit area and \( V \) the reverse bias applied to a Schottky barrier placed upon the semiconductor. We choose the voltage sign such that \( v > 0 \), that is, as the potential of the \( (n-\text{type}) \) semiconductor relative to the metal. In the depletion approximation of \( pn \) junction and Schottky barrier theory, the quantity \( w \) is the depth of the edge of the depletion layer.

The average process that leads from \( n(x) \) to \( \hat{n}(x) \) was described conceptually by Johnson and Panousis (4) (JP). Quantitative results were obtained numerically, by assuming certain true \( n(x) \) properties, and then simulating the \( C-V \) profiling process on a computer, obtaining the apparent electron concentration profile \( \hat{n}(x) \) through simulation. While tedious, the process is straightforward, and it appears to have been widely used since the work of JP. Unfortunately, the problem most commonly encountered in practice is the inverse one: to obtain \( n(x) \) from \( \hat{n}(x) \).

This problem is rather more difficult. In practice, it has meant trial and error. One guesses a plausible \( n(x) \) profile, simulates the \( C-V \) process to obtain \( \hat{n}(x) \), and compares the simulated with the measured profile. Based on the remaining discrepancies, one could guess a better \( n(x) \) profile, and repeat the process, but the convergence properties of such an iteration are uncertain at best, and such an iteration has rarely been used.

In practice, this problem has proven to be a more severe limitation on the accuracy of \( C-V \) profiling than the distinction between \( n(x) \) and \( n_d(x) \), particularly in the study of ion implanted layers, and of isotope heterojunctions. In fact, JP pointed out that the KOB correction is of little use because of this circumstance.

In this paper we will show that the \( n \rightarrow \hat{n} \) – Debye spreading process conserves not only the total charge contained in the true electron distribution- which is probably obvious -but also the moment of this distribution.

\[ \int_0^n \hat{n}(x) x \, dx = \int_0^n n(x) x \, dx . \]  \eqref{3}

Here \( x = 0 \) is the Schottky barrier plane, and the upper integration limit \( x = \infty \) is to be understood symbolically. We assume that deep inside the semiconductor, \( n(x) \) will eventually become flat, so that \( \hat{n}(x) \rightarrow n(x) \rightarrow n_c(x) \).

Equation \eqref{3} holds if the integration is extended into his range. This \textit{moment conservation theorem} –a generalization of a simple theorem derived by one of us (5) in 1956- has proven to itself a valuable aid in reconstructing from measured nonuniform apparent \( n(x) \) profiles certain important quantitative details about the non-uniformity of the underlying true \( n(x) \) profiles. In a recent paper (6) we had already applied the theorem to estimate the conduction band discontinuity in
GaAs–(Al,Ga) As \( n - N \) heterojunctions from \( C - V \) profiles taken through these junctions. In the present paper we give a proof of the theorem and show how it may be used to reconstruct the true doping gradient at a graded doping step from the more heavily graded apparent profile. Finally, in Section 4, we give a more quantitative measure for what is meant by the statement that the apparent electron distribution is an average of the true \( n(x) \), “taken over a distance of the order of a Debye length”.

### 2. Conservation of Charge and Moment

Let \( n(x) \) be the true electron concentration inside a semiconductor, in the presence of a reverse bias \( V \) applied to a Schottky barrier placed upon the surface of the semiconductor at \( x = 0 \). Suppose now that the reverse bias is increased by a small voltage \( \Delta V \), to \( V + \Delta V \). This will deplete an electron distribution \( \Delta n(x) \) from the semiconductor, so that the new electron distribution is \( n(x) - \Delta n(x) \). We call \( \Delta n(x) \) the incremental displaced electron distribution (IDED) (Fig. 1). The total capacitive charge \( \Delta Q \) flowing as a result of the voltage change \( \Delta V \) is

\[
\Delta Q = q \Delta N, \quad \text{where} \quad \Delta N = \int_0^{\infty} \Delta n(x) dx.
\]  

As a result of the depletion of electrons, the magnitude of the electric field inside the semiconductor changes by

\[
\Delta E(x) = \frac{q}{\varepsilon} \int_0^{\infty} \Delta n(y) dy.
\]  

The voltage increment \( \Delta V \) follows by a second integration:

\[
\Delta V = \int_0^{\infty} \Delta E(x) dx = \frac{q}{\varepsilon} \int_0^{\infty} dx \int_0^{\infty} \Delta n(y) dy,
\]  

where we have chosen the sign such that a reverse bias corresponds to a positive voltage. Integration by parts yields

\[
\Delta V = + \frac{q}{\varepsilon} \int_0^{\infty} \Delta n(x) dx,
\]

that is, the incremental reverse bias is simply \( q/\varepsilon \) times the moment of the IDED.

The Schottky barrier differential capacitance per unit area is, by definition,

\[
C = \frac{\Delta Q}{\Delta V} = \frac{1}{\varepsilon} \int_0^{\infty} \Delta n(x) dx / \int_0^{\infty} \Delta n(x) dx = \frac{\varepsilon}{(x)}.
\]

Here

\[
(x) = \frac{1}{\Delta N} \int_0^{\infty} \Delta n(x) dx
\]

is the mean location of the IDED. A measurement of the differential capacitance per unit area is therefore a measurement of the mean position of the differential charge. The result is exact.

We may view \( C - V \) profiling as the successive application of small voltage step is a distribution \( \Delta n_i(x) \) of depleted electrons with a mean position \( \langle x \rangle_i \) and a total number of depleted electrons \( \Delta N_i \). It is convenient to assume that the size of successive voltage steps are such that successive values of \( \langle x \rangle_i \) are incremented by the same amount \( \Delta x \).
\[ \langle x \rangle_{i+1} - \langle x \rangle_i = \Delta x \text{ independant of } i \]  

(10)

The important point is now that, for sufficiently small values of \( \Delta x \), the successive distributions \( \Delta n_i(x) \) will overlap strongly: the width of each \( \Delta n_i(x) \) will be much larger than the amount \( \Delta x \) by which the mean positions advance (Fig. 2a).

It is this overlap that is ignored when \( C-V \) data are interpreted using the depletion approximation, which is implied in equation (2). In the depletion approximation, the gradual falloff of the true electron concentration near the edge of the depletion layer is replaced by an abrupt step, which advances with increasing reverse bias. In effect, for a sufficiently small voltage increment, the true depleted electron distribution is replaced by a rectangular strip of width \( \Delta x \), with a suitable height \( n \), (Fig. 2b). In order that this interpretation preserves the experimental \( C-V \) relationship, it is
necessary to choose $\hat{n}$, and the location of the strip such that each strip corresponds to the same incremental voltage as the true incremental distribution $\Delta n_i(x)$. Equality of charge

$$\hat{n}_i \Delta x = \int_0^{\infty} \Delta n_i(x) dx.$$ (11)

Equality of voltage increments requires that the moment of the strip is the same as the moment of the true $\Delta n_i(x)$. Taken together with (11) this simply means that the center of the strip must coincide with the mean position $\langle x \rangle_i$ of the electrons in $\Delta n_i(x)$.

These requirements fully specify the value of $\hat{n}$, associated with each voltage increment, as well as the $x$-position to which the value is assigned. The result is a step-like apparent electron distribution $\hat{n}_i(x)$ of the form shown in Fig. 2b. Going to the limit $\Delta x \to 0$ leads to a continuous curve, $\hat{n}(x)$.

If the true electron concentration is uniform over at least as wide as the width of the incremental distribution $\Delta n_i(x)$, then the construction principle for $\hat{n}(x)$ leads to the same value as $n(x)$. Consider now a true electron concentration distribution in which a region of varying $n(x)$ is embedded between two sufficiently wide regions of uniform $n(x)$ (Fig. 3). Let $(a,b)$ be an interval whose end points fall sufficiently far into the two regions of uniform $n(x)$ that

$$\hat{n}(a) = n(a) \quad \text{and} \quad \hat{n}(b) = n(b).$$

Suppose that we are profiling from $a$ to $b$, by applying a sequence of small voltage steps, leading to a strip-like apparent electron concentration $\hat{n}(x)$, as described above. The height and the location of each individual strip are such that both the charge contained in each strip and the voltage contributed by it are equal to the true charge and the true voltage increment in each corresponding voltage step. The same must be true for the sum of all charges and the sum of all voltage contributions. But this implies conservation of total charge.

$$\int_a^b \hat{n}(x) \ dx = \int_a^b n(x) \ dx.$$ (12)

and conservation of total moment.

$$\int_a^b \hat{n}(x)x \ dx = \int_a^b n(x)x \ dx.$$ (13)

in profiling from $a$ to $b$. Fig. 3 Example of a graded doping step $n_d(x)$, together with its associated true and apparent electron distributions $n(x)$ and $\hat{n}(x)$. The $n(x)$ and $\hat{n}(x)$ curves must cross each other at least twice to conserve both area and moment. The curves shown are qualitative only; the difference between $n(x)$ and $\hat{n}(x)$ has been exaggerated, for clarity.
3. Application to Graded Doping Steps

It is evident that the moment conservation theorem can aid in reconstructing the true electron concentration from the apparent electron concentration: Any correction applied to the measured apparent electron distribution must not only have zero net charge, but also zero net moment, thus greatly restricting the range of possible corrections in a trial-and-error reconstruction. For example, any correction must at least have two nulls, that is, there must be at least two crossovers of the true and the apparent electron concentration. In the work of JP (4), this is evident only in Fig. 6 of their paper: in the other cases one of the $n \rightarrow n$ crossovers always lies just at the edge or even outside the range plotted. In the work of Kroemer et al (6), there are three and four crossovers, due to the greater complexity of the profiles.

Less obvious than the utility of moment conservation in reconstructing the true $n(x)$ profile, but probably of greater practical importance, is that under favorable circumstance one may be able to extract essential information about the semiconductor from the magnitude of the measured moment itself, without having to undertake the more difficult task of reconstructing the entire true electron profile. This was first demonstrated by Kroemer et al, [6] in their study of GaAs – (Al,Ga) as $n \rightarrow N$ heterojunctions, where it was possible to estimate the magnitude of the conduction band discontinuity from the magnitude of the moment of the apparent electron distribution. This was first demonstrated by Kroemer et al, [6] in their study of GaAs – (Al,Ga) as $n \rightarrow N$ heterojunctions, where it was possible to estimate the magnitude of the conduction band discontinuity from the magnitude of the moment of the apparent electron distribution.

We apply here the moment conservation theorem to a case of perhaps more widespread interest: The estimation of the degree of grading of a doping step, as it might occur at an epitaxial substrate interface. We assume a doping profile as shown in Fig. 3 with sufficiently wide flat doping regions on either side of the step so that in those regions the apparent electron profile coincides with both the true electron profile and the donor concentration:

$$\hat{n}(a) = n(a) = n_d(a), \quad \hat{n}(b) = n_d(b).$$  \(14\)

Associated with such a step is a built-in voltage

$$V_{bi} = (kT/\epsilon) \ln \left[ \frac{\hat{n}(b)}{\hat{n}(a)} \right].$$  \(15\)

Here we have assumed a non-degeneracy; if one or both of the terminal doping levels is degenerate, the Joyce-Dixon approximation (7) should be used to apply appropriate corrections.

On the other hand, the built-in voltage must be related to the net dipole moment formed by the deviation of the electron distribution from the donor distribution:

$$V_{bi} = (q/\epsilon) \int_a^b \left[ n_d(x) - n_d(x) \right] dx.\ \ (16)$$

Electrical neutrality, and the conservation of charge and moment require

$$\int_a^b n_d(x) dx = \int_a^b \hat{n}(x) dx,\ \ (17)$$

$$\int_a^b n_d(x)x dx = \int_a^b \hat{n}(x)x dx - (\epsilon kTq/\epsilon) \ln \left[ \hat{n}(b)/\hat{n}(a) \right]. \ \ (18)$$
Hence an accurate measurement of \( \hat{n}(x) \) gives accurate values for both the total donors contained in the interval \((a,b)\) and for the moment of their distribution.

Now there will be reasons to assume that the shape of the gradient satisfies a known mathematical model, be it something as simple as a linear gradient, as discussed in JP, or an error function distribution as for the simplest diffusion model, or something else. It may then be possible to determine the numerical values of the adjustable parameters of such a model, by fitting those parameters so that (17) and (18) are satisfied.

We show this here for an error function profile. We assume

\[
n_d(x) = \hat{n} - \frac{1}{2} \delta n \text{erf} \left( \frac{x}{2L} \right). \tag{19}\]

Here

\[
\frac{1}{2} \left[ n_d(a) + n_d(b) \right] n_d(a) - n_d(b). \tag{20}
\]

The length parameter \( L \) is the diffusion length, \( L = (Dt)^{1/2} \), and we have moved the coordinate origin to the center of the gradient. This implies that the coordinate origin of the experimental \( \hat{n}(x) \) profile is chosen such that

\[
\int_a^b \hat{n}(x) \, dx = (b-a) \hat{n}(a) - b \left[ \hat{n}(a) - n(b) \right]. \tag{21}
\]

Note that \( a < 0 \). For the following it is convenient to re-write (19) in the form

\[
n_d(x) = n^*(x) \pm \frac{1}{2} \delta \text{erfc} \left( \frac{x}{2L} \right), \text{ for } x > 0. \tag{22}\]

where \( n^*(x) \) would be the limit of a perfectly abrupt distribution,

\[
n^*(x) = \begin{cases} n_d(a) & \text{for } x < 0 \\ n_d(b) & \text{for } x > 0 \end{cases} \tag{23}\]

We insert the form (22) on the l.h.s. of (18) and assume that \(-a >> L, b >> L\). In this limit

\[
\int_a^b n_d(x) \, dx \equiv \int_a^b n^*(x) \, dx + \delta n \int_a^b \text{erfc} \left( \frac{x}{2L} \right) \, dx
\]

\[
= \int_a^b n^*(x) \, dx
\]

\[
+ 4L^2 \delta n \int_0^\infty \text{erfc} \left( s \right) ds. \tag{24}\]

The integral over the error function has the value \( \frac{1}{4} \). If this is inserted into (24) and the result into (18), the latter expression may be re-arranged to read

\[
L\delta n = \int \left[ \hat{n}(x) - n^*(x) \right] \, dx - (a, b, L) \ln \left( \frac{\hat{n}(b)}{\hat{n}(a)} \right). \tag{25}\]

The integral on the right is simply the moment of the deviation of the measured apparent electron distribution from an ideal abrupt distribution with the same total charge. Equation (25) expresses the unknown quantity \( L \) in terms of quantities all of which can be determined directly from the measured \( \hat{n}(x) \) profile.

The accuracy of the method is limited by the accuracy of the experimental data. The two terms on the r.h.s. of (25) have opposite signs, and for an abrupt doping transition they cancel exactly. Very steep doping gradients, for which the
difference between two terms in (25) is no larger than the experimental uncertainty in their magnitude, can clearly not be determined this way.

It should be evident that the method described here is easily applied to mathematical models other than an error function distribution.
the number of thermal equilibrium electrons arriving at the interface is not sufficient. We therefore proceed next to the calculation of this correct number.

5.2.2) The Richardson Equation.

We treat electrons as particles in a cube–shaped box of linear size \( L \), and volume \( V = L^3 \). Quantum–mechanically, the electrons are plane waves with a wave vector \( \mathbf{k} \). Periodic boundary conditions demand that for each of the components of \( \mathbf{k} \)

\[
k_L = 2\pi \mathcal{N} ,
\]

where \( \mathcal{N} \) is an integer. The separation between adjacent \( k \)'s is therefore

\[
\Delta k = \frac{2\pi}{L} .
\]

The kinetic energy of the electrons is

\[
E = E_x + E_y + E_z ,
\]

where we have, for the contribution due to motion in the \( x \)-direction,

\[
E_x = \frac{\hbar^2 k_x^2}{2m^*} ,
\]

with analogous relations for \( E_y \) and \( E_z \). At total energies sufficiently far above the Fermi level, the total number of electrons in a volume element \((dk)^3\) of \( k \)-space is

\[
dN = 2 \cdot \exp \left[ \frac{E - E_F}{kT} \right] \frac{dk_x}{(\Delta k)^3} \frac{dk_y}{(\Delta k)^3} \frac{dk_z}{(\Delta k)^3} .
\]

Here, the leading factor 2 is due to spin. The current density contributed by the electrons in (15) is simply

\[
dJ_z = -q \cdot \mathbf{v}_z dN/L^3
\]

if \( k_z > 0 \) and \( E_z > E_B \), and zero otherwise. Integration over all \( \mathbf{k} \)-values that satisfy these conditions yields, after some manipulation,

\[
J_z = \frac{2q}{(2\pi)^3} \cdot I_x \cdot I_y \cdot I_z \cdot \exp \left[ + \frac{E_F}{kT} \right] ,
\]

where \( I_x, I_y, \) and \( I_z \) are three integrals, given by
\[ I_x = I_y = \int_{-\infty}^{+\infty} \exp \left[ \frac{\hbar^2 k_x^2}{2m^*kT} \right] dk_x = \frac{\sqrt{2\pi m^*kT}}{\hbar}, \quad (18) \]

\[ I_z = \int_{k_0}^{\infty} \exp \left[ \frac{-\hbar^2 k_z^2}{2m^*kT} \right] \frac{\hbar k_z}{m^*} dk_z = \frac{kT}{\hbar} \exp \left[ \frac{E_B}{kT} \right], \quad (19) \]

In the last integral, the lower limit \( k_0 \) and the barrier height \( E_B \) are related via

\[ E_B = \frac{\hbar^2 k_0^2}{2m^*}. \quad (20) \]

Putting it all together, we obtain the famous Richardson Equation

\[ J = -A T^2 \exp \left[ -\frac{q\Phi_B}{kT} \right], \quad (21) \]

where

\[ A = \frac{4\pi q m^* k^2}{(2\pi \hbar)^3} \approx 120 \frac{A}{cm^2K^2} \frac{m^*}{m}, \quad (22) \]

is the effective Richardson constant for a semiconductor with effective mass \( m^* \), and

\[ q\Phi_B = E_B - E_F, \quad (23) \]

is the Schottky barrier height, measured from the Fermi level of the metal.

5.2.3) Arrhenius Plots and their Pitfalls.

The barrier height \( q\Phi_B \) may be determined from experimental current density–vs. temperature data, by plotting the logarithm of the ratio \( J/T^2 \) as a function of \( 1/T \). According to (22), such a plot, called an Arrhenius plot of \( J/T^2 \), should yield a straight line with a slope \( q\Phi_B/k \), and an intercept with the vertical axis at the logarithm of the constant \( A \) in (22). If one does this, one often finds a quite different intercept. The principal reason for the deviation is that the barrier height itself tends to be slightly temperature–dependent. Suppose we have a linear dependence of \( q\Phi_B \) over the temperature range of the data,
q\Phi_B = q\Phi_{B0} - \alpha \cdot kT. \quad \text{(25)}

Insertion of this form into (22) yields

\[ J = -Ae^{\alpha/T^2} \exp \left( \frac{-q\Phi_{B0}}{kT} \right). \quad \text{(26)} \]

This is of the same form as (22), except that the constant A has been replaced by $Ae^{\alpha/T}$, and the slope of the Arrhenius plot is not given by the actual barrier height anywhere within the temperature range of the data, but by the value linearly extrapolated to $T = 0$. Note that this value may differ significantly from the actual barrier height at $T = 0$, because the temperature dependence tends to become weaker as $T \to 0$.

Similar precautions apply to all forms of Arrhenius plots of quantities whose temperature dependence is dominated by a Boltzmann factor.

### 5.3) Image Force Barrier Lowering at Schottky Barriers.

The barrier height of Schottky barriers depends significantly on the electric field strength $E$ at the tip of the barrier, due to the effect of the electrostatic image force. When an electron is placed a distance $x$ ahead of a metal surface, an image charge is induced on the metal surface, which attracts the electron towards the metal, with the image force. It is an exercise in elementary electrostatics to show that the image force is given by

\[ F = \frac{q^2}{16\pi \varepsilon x^2}, \quad \text{(31)} \]

where $\varepsilon$ is the permittivity of the semiconductor. This force corresponds to an attractive potential energy, relative to the potential energy at infinity, of

\[ E(x) = \frac{q^2}{16\pi \varepsilon x}. \quad \text{(32)} \]

To this potential energy we must add the true electrostatic potential energy $-q|E|x$ (Fig.1).

The sum of the two potential energies has a maximum at the position
\[ x = \sqrt{\frac{q}{16\pi \varepsilon \varepsilon_0}} \]  

which is typically of the order \(10^{-7}\) cm. At the potential maximum, the potential energy is lowered relative to the potential without image force by the amount

\[ \Delta \Phi_B = \sqrt{\frac{q\varepsilon_0}{4\pi \varepsilon \varepsilon_0}}. \]  

Fig. 1 Potential energy diagram ahead of a metal surface. The metal work function is \(q\Phi_m\). The potential barrier is lowered when an electric field is applied to the surface, due to the combined effects of the field and the image force.

For an electric field of, say 100 kV/cm, and a dielectric constant of 10, one finds 37.95 mV, implying the scaling relation

\[ \Delta \Phi_B = 37.95 \text{ mV} \cdot \sqrt{\frac{10\varepsilon_0}{\varepsilon}} \cdot \frac{\varepsilon}{100 \text{ kV/cm}}. \]  

Evidently, with increasing reverse bias voltage the electric field at the interface increases, which pulls the barrier down (Figure 6.3–2), thereby increasing the reverse current.
Because the distance of the potential maximum from the metal is so small, an electron with a typical thermal velocity of about $10^7 \text{cm/s}$, the electron traverses the barrier in a very short time, of the order $10^{-14} \text{sec}$. The permittivity $\varepsilon$ to be used in (34) and (35) is therefore not the static permittivity, but the high-frequency permittivity for frequencies of the order $10^{14} \text{Hz}$. These are infrared frequencies, corresponding to wavelengths of about $3 \mu \text{m}$. However, for most semiconductors, the infrared permittivities remain close to the static permittivities up to the energy gap equivalent frequency $E_g / h$, which is usually larger than $10^{-14} \text{sec}$. Hence, the dynamic corrections to the image force tend to be small. Fig. 3 gives an example, for Si.

![Energy-band diagram](image)

**Fig. 2.** Energy-band diagram incorporating image force barrier lowering under different biasing conditions. The intrinsic barrier height is $q\Phi_{B0}$. The barrier height at thermal equilibrium is $q\Phi_{Bn}$. The barrier lowerings under forward and reverse bias are $\Delta \Phi_F$ and $\Delta \Phi_R$, respectively. (After Rideout)
We consider electron flow in only one direction, parallel or anti-parallel to the electric field $\mathcal{E}$. We assume that, whenever the electron reaches a certain maximal speed $v_m$, it is immediately scattered, losing a certain amount $\Delta E$ of kinetic energy in the process. As a result, its speed will drop to a value $v_r < v_m$. Its velocity after the scattering may be either $+v_r$, for a forward scattering event, or $-v_r$, for backward scattering. Inelastic scattering processes tend to be isotropic, so we make the assumption that forward and backward scattering are equally probable. We are only interested in the time-averaged electron velocity; for equal scattering probabilities it may be calculated as if forward and backward scattering events alternated, rather than being randomly distributed. For alternating forward and backward scattering, the electron velocity as a function of time behaves as shown in Fig. 3.

![Velocity vs. time for electrons in the saturated drift velocity regime](image)

**Fig. 3:** Velocity vs. time for electrons in the saturated drift velocity regime

Following each scattering event, the electron is accelerated from the velocity $\pm v_r$ to the velocity $v_m$, at which point the next scattering event takes place. Following a forward scattering event, the electron reaches the velocity after a time interval

$$t_+ = (v_m - v_r) \cdot \frac{m^*}{q \mathcal{E}}.$$  (**13-7a**)

If no other kinds of scattering events intervene. Following a backward scattering event, the electron is first decelerated, and then accelerated again, reaching the velocity $v_m$ after a time...
\( t_\perp = (v_m + v_r) \cdot m^*/qE \).  \hspace{1cm} (13-7b)

again assuming that no other kinds of scattering events intervene.

The qualifiers with regard to other kinds of scattering events are important. But we note from (13-7a,b) that the times to the next inelastic scattering event decrease with increasing field strength, thereby decreasing the probability of another kind of scattering event intervening. At sufficiently large fields, such alternate events become negligible. The average velocity in that limit, called the saturated drift velocity, is then easily found to be

\[ v_s = (v_m^2 - v_r^2)/2v_m, \]  \hspace{1cm} (13-8)

independent of the electric field strength! Although the rate of acceleration increases with increasing field strength, the rate of velocity-reducing collisions increases in the same way, and the field strength cancels out of the average velocity.

We is often useful write (13-8) in an alternate way,

\[ v_s = v_m \Delta E/2E_m, \]  \hspace{1cm} (13-9)

where \( E_m \) is the kinetic energy associated with the velocity \( v_m \).

Although highly oversimplified, our treatment captures the essence of the physics of drift velocity saturation. It is readily possible to refine the model, but the mathematical effort soon becomes prohibitive, without a major change in the final result.

In our treatment, we neglected the near-elastic scattering that dominates the electron dynamics at low electric fields; hence this treatment describes only the high-field limit. At low fields, we have, of course the simple linear relation

\[ v_D = \mu_e E. \]  \hspace{1cm} (13-10)

Somewhere in the vicinity of the critical field

\[ E = E_c = v_s/\mu_e \]  \hspace{1cm} (13-11)

the two relationships cross over. A mathematical theory that would cover the crossover range as well would be a difficult undertaking. For purposes of device physics calculations it is usually sufficient to somehow interpolate between the two relations. Two simple interpolations are widely used:
\[ t^+ = \frac{(V_m - V_r)}{q \bar{V} \bar{E}} \cdot \frac{m^*}{q \bar{V} \bar{E}} \]

\[ t^- = \frac{(V_m + V_r)}{q \bar{V} \bar{E}} \cdot \frac{m^*}{q \bar{V} \bar{E}} \]

\[ t_0 = \frac{V_r \cdot m^*}{q \bar{V} \bar{E}} - V_r \cdot a, \quad t^- t_0 = V_m \cdot a \]

\[ V_{aw} = \frac{1}{T} \left[ -V_r \cdot \frac{t_0}{2} + V_m \left( \frac{t^- - t_0}{2} \right) + \left( V_m - V_r \right) \cdot \frac{t^+}{2} + V_r \cdot t^+ \right] \]

\[ = \frac{1}{T} \left[ -\frac{V_r^2}{2} \cdot a + \frac{V_m^2}{2} \cdot a + \frac{(V_m - V_r)^2}{2} \cdot a + V_r \cdot (V_m - V_{aw}) \right] \]
\[ V_{av} = \frac{q}{2T} \left[ \left( -V_r^2 + V_m^2 + V_m + V_r \right) + \left( V_m - V_r \right)^2 \right] \]
\[ = \frac{q}{T} \left[ V_m - V_r \right] \]
\[ = \frac{q}{T} \cdot (V_m - V_r) \]
\[ V_{av} = \frac{V_m^2 - V_r^2}{2V_m} = V_S \]
\[ \Delta E = \frac{1}{2} m^* \left[ V_m^2 - V_r^2 \right] \]
\[ \Rightarrow V_S = \frac{\Delta E}{m^* V_m} = \frac{V_m \cdot \Delta E}{2E_m} \]
(a) Assume a piecewise linear $v(\mathcal{E})$ characteristic: One simply uses (13-10) for fields below $\mathcal{E}_c$ and the saturated drift velocity $v_s$ above. Although crude, such an approximation captures the essence of having both a linear and a saturated drift velocity range, and it often makes problems mathematically tractable that would be intractable under more realistic assumptions.

(b) Assume a continuous interpolation of the form

$$\frac{1}{V_D} = \frac{1}{\mu_e \mathcal{E}} + \frac{1}{v_s}.$$  

(*13-12)

For low fields the first term is dominant, thus yielding the linear limit (13-10), for high fields the second term dominates, leading to saturation of the drift velocity. Because of the absence of a mathematical singularity, the relation (*13-12) is sometimes easier to use than the piecewise linear model. The actual $v(\mathcal{E})$ characteristics of semiconductors tend to saturate more abruptly than described by (*13-12), falling somewhere between (13-12) and the piecewise linear model.

### 13.4) Space Charge Effects in the Base-Collector Depletion Layer (Kirk Effect)

One of the consequences of drift velocity saturation is that it is not possible to speed up the current transport through the base-to-collector depletion layer of a bipolar transistor by simply increasing the collector reverse bias. At high current densities that leads to appreciable space charge effects inside the depletion layer, and there will be a certain maximal current density beyond which the collector ceases to collect carriers effectively.

We assume that the base/collector junction is a $p^+n^-n^+$ junction, with an $n^-$ region of width $w$. In the absence of any current, the electric field distribution within the depletion layer will be as in line "a" in Fig. 4, with the maximal field occurring at the base end of the depletion layer: With increasing current this distribution first flattens out and then reverses; for constant reverse bias, the area under the line remains constant, too. At a current density

$$J_C = qn^-v_s$$  

(*13-13)

the space charge of the traveling current exactly cancels that of the stationary donors, and the field becomes uniform (line "b"). For higher current densities the field distribution reverses. Eventually, the electric field at the base end of the depletion
region will go to zero (line "c"); this happens when the current density reaches the value

\[ J_C = J_{\text{max}} = \left[ \frac{2eV_C R}{w^2} + qn_- \right] v_s. \quad (\text{13-14}) \]

*Fig. 4: Electric Field distribution in the collector depletion layer of a transistor, for various current densities.*

When the current density exceeds this value, the overall result is the a drastic extension of the effective base width into the region between the metallurgical base and the collector (Kirk Effect; see Sze, pp 145 ff.). See *Fig. 5* on next page.
TWO INTEGRAL RELATIONS PERTAINING TO THE ELECRO TranS Port THROUGH A BllP TranS TrANSP ORR WITH A NOnUNIFORM ENERgy GAP IN THE BASE REGION

H. KROEMER

Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106, U.S.A.

(Received 23 November 1984; in revised form 23 February 1985)

Abstract—The two integral relations by Moll and Ross for the current flow through the base region of a bipolar transistor, and for the base transit time, are generalized to the case of a heterostructure bipolar transistor with a nonuniform energy gap in the base region.

The recent strong interest in heterostructure bipolar transistors (HBTs) with a nonuniform energy gap in the base region[1] calls for an extension of bipolar transistor theory to transistors that exhibit such a variation. The first priority is a generalization of the familiar integral relation for the current flow through the base region,

\[ J_n = -\frac{qDnn^2}{N_b} \exp \left( \frac{qV_{BE}}{kT} \right) \int N_a \, dx \]  \hspace{1cm} (1)

This relation was first derived in slightly different form by Moll and Ross[2]; it forms the backbone of the Gummel–Poon model[3] widely used in transistor modeling, for example in the SPICE computer program[4].

According to (1), the current density for a given base/emitter voltage depends only on the total number of acceptors per unit area in the base, the so-called Gummel Number, but not on the spatial distribution of these acceptors. As we shall see, this changes drastically in a graded-gap structure.

A second quantity of interest is the electron transit time through the base region. It does depend on the acceptor distribution; Moll and Ross[2] first derived the relation

\[ \tau = \frac{1}{D_{hT} \int N_b} \int \left( \int \frac{N_a \, dx}{N_b} \right) \, d\tau \]  \hspace{1cm} (2)

A very “physical” rederivation and discussion of this relation was subsequently given by Varner[6].

The relations (1) and (2) are for the simplest case, when a number of approximations may be made, such as low-level injection, position-independent diffusivity, negligible hole current, and absence of hot-electron effects causing nonlinearity in the basic drift-and-diffusion transport relations. Refinements of the current flow relation (1) that remove the various approximations, but which retain the uniform energy gap, have been extensively discussed in the literature[3–5]. The transit time relation (2), however, has been largely ignored by subsequent workers[7].

Our point of departure is the set of linear transport relations for the electron and hole contribution to the electrical current density across the base region of an npn-transistor, expressed in their most general form, involving the two carrier concentrations and the gradients of the two quasi-fermi levels \( E_n \) and \( E_p \) for electrons and holes:

\[ J_n = +q\mu_n n \frac{d}{dx} \left( \frac{E_n}{q} \right) = +\mu_n n \frac{d}{dx} E_n, \]  \hspace{1cm} (3a)

\[ J_p = -q\mu_p p \frac{d}{dx} \left( \frac{E_p}{q} \right) = -\mu_p p \frac{d}{dx} E_p. \]  \hspace{1cm} (3b)

This formulation implies the neglect of hot-electron effects. In a good transistor, the hole current may also be neglected, which implies \( dE_p/dx = 0 \). Because HBTs are likely to have base doping levels much higher than homostructure transistors, and higher \( \beta \), this is an even better approximation for HBTs than for the homostructure transistor. It permits rewriting (3a) in the form

\[ J_n = \mu_n n \frac{d}{dx} (E_n - E_p). \]  \hspace{1cm} (4)

Now, for nondegenerate densities,

\[ np = n_i^2 \times \exp \left[ \frac{(E_n - E_p)}{kT} \right], \]  \hspace{1cm} (5)

from which

\[ \frac{d}{dx} (E_n - E_p) = kT \times \left[ \frac{n_i^2}{np} \right] \times \frac{d}{dx} \left[ np/n_i^2 \right]. \]  \hspace{1cm} (6)
Insertion into (4) and rearrangement yields

$$\frac{d}{dx} \left( \frac{np}{n_i^2} \right) = \left( J_n/q \right) \times \left( \frac{p}{D_a n_i^2} \right),$$

(7)

where we have also used the Einstein relation \( qD_a = \mu_e kT \). Integration of (7) from an arbitrary point \( x \) within the base region to the collector edge \( x = w \) yields

$$\left. \frac{np}{n_i^2} \right|_w - \left. \frac{np}{n_i^2} \right|_i = \left. \frac{-np}{n_i^2} \right|_i = J_n/q \times \int_c^w \frac{p}{D_a n_i^2} \, dy. $$

(8)

In the first equality we have assumed that the collector is reverse-biased, and that \( np \mid_i \) may therefore be neglected. At the emitter end of the base, \( x = 0 \), we apply Shockley boundary conditions:

$$np/n_i^2 \big|_0 = \exp \left( -qV_{BE}/kT \right).$$

(9)

If (9) is inserted into (8), and the result solved for \( J_n \), we obtain the desired generalization of the Moll–Ross current relation:

$$J_n = - \frac{q \times \exp \left( qV_{BE}/kT \right)}{\int_g^b \left[ \frac{p}{D_a n_i^2} \right] \, dx}. $$

(10)

We return to (8) for arbitrary \( x \), and write it in the form

$$n(x) = -(J_n/q) \times (n_i^2/p) \times \int_x^w \left( \frac{p}{D_a n_i^2} \right) \, dy. $$

(11)

Integration over the base and division by \(- J_n/q\) yields the average transit time through the base:

$$- \left[ q \int_0^w n(x) \, dx \right]/J_n \quad \Rightarrow \quad \tau = \int_0^w \left[ \left( \frac{n_i^2}{p} \right) \int_x^w \left( \frac{p}{D_a n_i^2} \right) \, dy \right] \, dz. $$

(12)

This is the desired generalization of the transit time relation (2). In the uniform-base case, the new relations (10) and (12) both reduce to their already known limits (1) and (2), as they must.

The relations (10) and (12) are the two central results of this paper.

As stated earlier, HBTs are likely to have base doping levels much higher than homostucture transistors. In fact, in those HBTs that are of practical interest, the doping levels in the base will be higher than in the emitter. As a result, the various kinds of high-level injection effects in which the electron concentration in the base rises significantly above the doping concentration, and which play such an important role in homostructure transistor theory, cannot occur in HBTs of practical interest, and may therefore be ignored. Hence, it will, as a rule, be an excellent approximation to replace the hole concentration \( p \) in (10) and (12) by the acceptor doping concentration \( N_A \), even at the highest current densities the emitter is capable of delivering.

It is evident from (10) and (12) that in a variable-gap transistor the acceptor doping concentration is not weighted uniformly throughout the base, as in the uniform-gap case, but with the weighting factor \( 1/n_i^2 \), which in turn is proportional to \( \exp (E_g/kT) \).

To illustrate the application of the new relations, especially of the transit time relation (12), it is useful to discuss, as an instructive example, the case of a linear variation of the energy gap with position,

$$E_g = E_{g0} - qFx = E_{g0} - \Delta E_g \left( x/w \right), $$

(13)

with all other parameters position-independent, and also assuming \( p = N_A \). Evidently, the quantity \( F \) in (13) is the built-in drift field for the electrons, and \( \Delta E_g \) is the total variation of the energy gap within the base region. Equation (13) implies

$$n_i^2 = \frac{n_{i0}^2 \times \exp (qFx/kT),}{\left( n_i^2/n_{i0}^2 \right)^2} \exp \left( qV_{BE}/kT \right), $$

(14)

where the notation \( n_i \) refers to the values of \( n_i \) taken at the position \( x \). If (14) is inserted into (10), the result may be written

$$J_n = \frac{-q^2 F D_a n_{i0}^2 \times \exp \left( qV_{BE}/kT \right)}{\left( n_i^2/n_{i0}^2 \right)^2}. $$

(15)

Note that, if emitter and collector are interchanged, the magnitude of the current remains unchanged; only the sign changes. This is not the case for the transit time: insertion of (14) into (12) yields, after some manipulation,

$$\tau = \left( w/\mu_e F \right) \times \left( 1 - (kT/qFw) \times \left[ 1 - \left( n_{i0}/n_{i\infty} \right) \right] \right). $$

(16)

For normal transistor operation we have \( n_{i0}/n_{i\infty} \ll 1 \), in which case (16) may be simplified to

$$\tau = \left( w/\mu_e F \right) \times \left( 1 - \left( kT/qFw \right) \right) \ll \tau_0, $$

(17)

where, on the second line, we have introduced, for comparison, the base transit time for a uniform-base transistor,

$$\tau_0 = w^2/2 D_a, $$

(18)

and where we have assumed that \( \Delta E_g \gg kT \).
For inverted operation the inequality \( n_{i0}/n_{i\infty} \gg 1 \) holds, and we find

\[
\tau = \left( \frac{w}{\mu_e F} \right) \times \left( 1 + \left( \frac{kT/qFw}{\tau_0} \right) n_{i0}/n_{i\infty} \right)^2
\]

\[
= 2\tau_0 \times \left( \frac{kT/\Delta E_k}{\tau_0} \right)^2 \times n_{i0}/n_{i\infty}^2 \gg \tau_0. \tag{19}
\]

The asymmetry in the forward and reverse transit times implied in (17) and (19) is self-evident. The results (16)–(19) for the specific case of a uniformly doped base with a linearly varying gap are identical to the expressions for a uniform-gap transistor in which a uniform electric field has been incorporated into the base region by means of an exponential doping profile, a case widely discussed during the 1950s\cite{8,2,6}. The simple limit (17) was recently restated by Hayes et al.\cite{1}. What is new is the generalization of (2) to essentially arbitrary variations of the energy gap, in addition to that of other base parameters.

Acknowledgements—I wish to express my thanks to Dr. Peter Asbeck for helpful discussions, and to Prof. Fred Linholm and Dr. James M. Early for helping me find some of the old literature on that subject.

REFERENCES


4. See, for example, L. Getreu, Modeling the Bipolar Transistor, Tektronix Inc. (1976).


EARLY VOLTAGE
(Physical picture)

\[ \frac{\Delta Q_B}{\Delta V_{CB}} = C_{jc} = \frac{\varepsilon}{W_c} V_A \]

\[ \Delta Q_B = \frac{\varepsilon}{W_c} \Delta V_{CB} \]

\[ Q_B = Q_{BC} - \frac{\varepsilon}{W_c} \Delta V_{CB} \]

\[ J_E = J_c(\Delta V_{CB}) = \frac{QDn_0(c)}{X_B} - \frac{Q'V_{CB}}{W_c} \]

\[ J_c = \frac{K}{X_{B_0} - \frac{\varepsilon}{W_c} \Delta V_{CB}} \]

When the denominator \( \to 0 \), \( J_c \to \infty \) Punchthrough

\[ J_c = \frac{K}{X_{B_0} - \frac{\varepsilon}{W_c} \Delta V_{CB}} \]

\[ \Delta J_c = \frac{K}{X_{B_0}} \frac{\Delta V_{CB}}{V_A} \]

\[ \Delta V_{CB} = \frac{V_A}{\Delta J_c} \]
specifically considered thus far. For example, a transistor as described in Chapter 6 would have a current-source output in the active mode; that is, output current would not depend in any way on the voltage applied between the base and collector terminals. In real transistors, however, both output currents and output voltages influence the device performance. Similarly, the theory that has thus far been presented would predict a current gain in the active mode that is insensitive to bias level. But this is only a rough approximation. A first objective of this chapter, therefore, is to investigate mechanisms that must be considered to obtain a more realistic picture of transistor operation.

Another major inadequacy of the theory developed thus far is that it does not deal with time-varying effects. Their consideration is therefore the second major topic to be developed in this chapter. A great simplification in considering time-varying effects will be obtained by modeling of transistors as charge-controlled devices. This viewpoint leads very naturally into the important topic of large-signal transient models for bipolar transistors. The amalgamation of the Ebers-Moll model with the charge-control model provides a powerful analysis technique. It will be straightforward to derive a small-signal ac equivalent circuit from the more general large-signal representation. This small-signal equivalent circuit, known as the hybrid-pi circuit, is relatively easy to characterize because it is composed of elements that relate directly to physical mechanisms in the transistor.

The production of pnp transistors within the confines of standard IC processing is described in a final section. Two basic types of pnp transistors, substrate pnp and lateral pnp transistors are typically fabricated. The substrate pnp transistor has only limited application because its collector is not isolated. The lateral pnp transistor has severely limited performance compared to that of npn transistors. Simple models for loss mechanisms in this device explain the limited performance. An important application of lateral pnp transistors—a dense form of logic circuitry known as integrated-injection logic (I^2L)—is described at the conclusion of the chapter.

### 7.1 Effects of Collector Bias Variation (Early Effect)

When we considered bipolar transistors under active bias in Chapter 6, the function of the voltage applied across the collector-base junction was merely to insure the efficient collection of base minority carriers and their delivery to the collector region. The magnitude of the bias only limited the range of the permissible collector voltage swing, provided that it was below the breakdown value. We have overlooked an important aspect of pn-junction electronics in taking this simplified view. As we noted in Chapter 4, the width of a reverse-biased pn junction is voltage dependent; in fact, this bias dependence made possible the operation of junction field-effect transistors. In the case of bipolar transistors, a changing collector-base bias causes variation in the space-charge layer width at the collector junction and, consequently, in the width of the quasi-neutral base region. These variations result
7.1 Effects of Collector Bias Variation (Early Effect)

In several effects that complicate the performance of the device as a linear amplifier. Base-width modulation as a consequence of variations in collector-base bias was first analyzed by James Early,1 and the phenomenon is generally called the Early effect.

The dependence of collector current on collector-base bias can be formulated directly by using the integral equations for active bias (nnp transistor) developed in Section 6.2. In particular, from Equations 6.2.1 and 6.1.10, we can write

$$I_C = \frac{q D_n n_0^2 A_k \exp(qV_{BB}/kT)}{\int_0^{x_B} p \, dx} \left( \frac{\partial I_C}{\partial V_{CB}} \right)$$  \hspace{1cm} (7.1.1)

where the integration is performed over \( x_B \), the width of the quasi-neutral base region and the other terms have been defined in Section 6.1.

Variation in the base width with voltage \( V_{CB} \) causes a variation in the collector current that can be written

$$\frac{\partial I_C}{\partial V_{CB}} = -\frac{q D_n n_0^2 A_k \exp(qV_{BB}/kT)p(x_B) \partial x_B}{\int_0^{x_B} p \, dx} \left( \frac{\partial x_B}{\partial V_{CB}} \right)$$  \hspace{1cm} (7.1.2)

Several of the terms in Equation 7.1.2 can be combined to represent collector current itself, so that Equation 7.1.2, which represents the small-signal conductance at the collector-base junction, can be written

$$\frac{\partial I_C}{\partial V_{CB}} = -I_C p(x_B) \left( \frac{1}{\int_0^{x_B} p \, dx} \right) \left( \frac{\partial x_B}{\partial V_{CB}} \right)$$

$$= -\frac{I_C}{V_A} = \frac{|I_C|}{V_A}$$  \hspace{1cm} (7.1.3)

Since the collector is reverse biased, the derivative in Equation 7.1.3, \( \partial x_B/\partial V_{CB} \), is negative, and the Early effect results in an increase in \( I_C \) when \( V_{CB} \) is increased. This increase is evident if we examine common-emitter output characteristic curves such as the family shown in Figure 7.1a. These curves are actually plots of \( I_C \) versus \( V_{CB} \), but \( V_{CE} \) has almost the same value as \( V_{CB} \) for bias in the active region \( (V_{CE} \approx V_{CB} + 0.7 \, V) \).

Equation 7.1.3 reveals that the Early effect varies linearly with collector current. The reciprocal of the factor multiplying the current has the dimension of a voltage and has been defined as the Early voltage. The Early voltage is usually given the symbol \( V_A \). From Equation 7.1.3, an equation for \( V_A \) for an nnp transistor is

$$V_A = \frac{\int_0^{x_B} p \, dx}{p(x_B)\partial x_B/\partial V_{CB}}$$  \hspace{1cm} (7.1.4)
Figure 7.1 Measured output characteristics for an amplifying transistor: collector current versus $V_{CE}$ (a) vertical 0.1 mA/div., horizontal 1 V/div. (b) vertical 0.1 mA/div., horizontal 10 V/div.; tangents to the measured curves (at the edge of saturation) are extended to the voltage axis (dashed lines) to determine the Early Voltage $V_A$. Extension of lines drawn approximately tangential to the characteristic in the active region intersects the voltage axis at $V_A'$ (solid lines).

Again, the derivative in Equation 7.1.4 is negative and, thus, the Early voltage is negative for an npn transistor. The analogous effect of emitter-base space-charge layer widening when the transistor is biased in the reverse-active mode can also be interpreted by using a different Early voltage, usually denoted as $V_B$.

Except for high-level effects, the three terms that define $V_A$ depend only on the transistor manufacturing process and on collector-base voltage. In practice, the collector-base voltage dependence of $V_A$ itself is usually treated as negligible and the Early voltage is approximated by its value at a single bias, often at $V_{CB} = 0$. By using this condition to specify $V_A$, we can expect that a series of tangents to curves of $I_C$ versus $V_{CB}$ ($V_{CE}$ in practice), drawn at the edge of the forward-active region where $V_{CB}$ is approximately zero, should intersect the $V_{CE}$ axis at $V_A$. In Figure 7.1b dashed lines drawn from the point at which the transistor moves out of saturation do intersect at a common point, indicating the Early voltage. For circuit-design and analysis, however, interest is not in an Early voltage characterizing the edge of saturation, but rather in a parameter to use in the forward-active region. If tangents are drawn to the curves of $I_C$ versus $V_{CE}$ in the active region, they do not, in general, intersect each other on the voltage axis. It is usual, however, to approximate an intersection point appropriate to the range of bias of the transistor as at $V_A'$, formed by the solid lines in Figure 7.1. This construction forms similar triangles $AOB$ and $BCD$ (if we neglect the small saturation-voltage offset of the BJT curves). We can infer Equation 7.1.3 directly from these similar triangles if the differential elements $\delta I_C$ and $\delta V_{CB}$ in Equation 7.1.3 are treated as incrementals $\Delta I_C$ (indicated by line $CD$) and $\Delta V_{CB}$ (approximately indicated by line $BC$) since line $AO$ indicates $V_A'$ and line $OB$ indicates $I_C$. 


7.1 Effects of Collector Bias Variation (Early Effect)

A useful and informative alternative expression for $V_A$ can be obtained by rearranging some terms in Equation 7.1.4. First, we use Equation 6.1.8 to express the numerator in terms of the base majority-charge density $Q_b$ in that portion of the base where transistor action is taking place:

$$\int_0^{x_B} p \, dx = \frac{Q_b}{q}$$  \hspace{1cm} (7.1.5)

We then recognize that the denominator of Equation 7.1.4 represents the derivative of base charge $Q_b$ with respect to $V_{CB}$:

$$\frac{dQ_b}{dV_{CB}} = \frac{dQ_b}{dx_B}$$  \hspace{1cm} (7.1.6)

The derivative on the right-hand side of Equation 7.1.6 can be related to $C_j$, the small-signal capacitance per unit area at the collector-base junction.

$$\left| \frac{dQ_b}{dV_{CB}} \right| = C_j$$  \hspace{1cm} (7.1.7)

Therefore, the Early voltage is

$$|V_A| = \frac{Q_b}{C_j}$$  \hspace{1cm} (7.1.8)

To reduce the influence of collector-base voltage on collector voltage, $V_A$ should be increased in magnitude. From Equation 7.1.8 we see that this can be accomplished in practice by increasing the ratio of base majority-charge per unit area to the capacitance per unit area at the collector-base junction. Physically, this reduces the movement of the base-collector boundary into the base region.

It is useful, for computer modeling purposes, to describe the widening of the base-collector space-charge layer through the Early voltage $V_A$, as we shall see in Section 7.6. It may, however, help conceptually to consider an alternative view of the Early effect that is physically more revealing for the prototype transistor introduced at the beginning of Chapter 6. For homogeneous doping and active-mode bias, the minority-carrier distribution in the base has a triangular shape (Figure 7.2). Increasing the collector-base voltage from $V_{CB}$ to $(V_{CB} + \Delta V_{CB})$ moves the base-collector junction edge a distance $\Delta x_B$ from $x_B$. A second triangular distribution specifies the new minority-carrier profile, and the shaded area between the two distributions represents the decrease in stored base charge.

The collector current is thus increased in the ratio $x_{B_0}/x_{B_1}$ by the change in $V_{CB}$. In addition, it becomes clear by looking directly at stored base charge that the Early effect reduces charge storage. This affects both the transient behavior of transistors and the dc base current since base recombination depends directly on stored base charge. We shall consider these effects in more detail in Section 7.5 when we discuss BJT models for circuit applications.
maintain acceptable injection efficiency [which constrains $N_d(0)$ to be appreciably lower than the emitter doping] and the requirement of an extrinsic $p$-type doped material at $x = x_n$ [which forces $N_d(x_n)$ to be greater than $N_{d,eq}$].

### 7.4 Charge-Control Model

A framework for bipolar-transistor equations that is especially useful for time-dependent analysis is called the \textit{charge-control model}.\(^7\) In this model the controlled variable is not current or voltage; instead, equations are framed in terms of controlled charges within regions of the device.

A typical charge-control relationship for a transistor under active bias was derived in the previous section in Equation 7.3.2. This equation, $I_C = \frac{Q_{nB}}{\tau_B}$ relates the minority charge stored in the quasi-neutral base $Q_{nB}$ to the current $I_C$ carried by transistor action between the emitter and the collector. The charge and current are linearly related with $\tau_B$, the transit time in the quasi-neutral base, as the proportionality factor. Because Equation 7.3.2 represents only minority-carrier transport across the base, however, it is just one portion of the charge-control model for a transistor.

Control in an amplifying $nnp$ transistor is exercised by the bias on the base-emitter junction. This bias affects not only $Q_{nB}$ but other charge components as well. The major additional components to be considered are the charges represented by holes injected into the emitter, which we designate as $Q_{pE}$, and the charges stored on the base-emitter and base-collector depletion capacitances, which are given the symbols $Q_{VE}$ and $Q_{VC}$, respectively. Figure 7.14 shows these components for a prototype transistor. We first discuss the two injection components, $Q_{nB}$ and $Q_{pE}$ that are responsible for steady-state base current. The other charge components shown in Figure 7.14 and will be considered later. The emitter voltage is increased when the $BJT$ is operating in the controlling (base majority) region.

Note that all that is formed in the quasi-neutral emitter region is similar to that of a diode, and we may write,

$$I_C = \frac{Q_{pE}}{\tau_B} = \frac{1}{\tau_B} \left[ \frac{Q_{pE}}{Q_{pE} + Q_{pE}} \right]$$

where $Q_{pE}$ is a function of $V_{BE}$ and $\tau_B$ is at least somewhat greater than $\tau_B$ (Equation 7.3.2). Further, the control model is developed later.

The steady-state current $I_E$ is the sum of the recombination processes for which $Q_{nB}$ recombines in the base, the injection processes for which $Q_{nB}$ injected into the emitter to the base, and the hole diode factor $[exp(qV_{BE}/kT) - 1]$ to $Q_{pE}$. It is therefore possible to obtain the collector (base) current of the transistor.

A considerable amount of effort is required to express $\tau_B$ usefully in the analysis is not of special value and is difficult to obtain from measurements, and a correct representation of the device for design is not possible.

Equations 7.4.1 and 7.4.2 state that the gain is simply the ratio of
models

7.4 Charge-Control Model

components shown in Figure 7.14 influence the time-varying behavior of the BJT and will be considered later. Since both \(Q_{nb}\) and \(Q_{pe}\) increase when the base-emitter voltage is increased, their sum \((Q_{nb} + Q_{pe})\) is called \(Q_F\) (because \(Q_F\) is increased when the BJT is under forward-active bias). The sign of \(Q_F\) is the sign of the controlling (base majority-carrier) charge, positive for an npn transistor and negative for a pnp transistor. The steady-state collector current can be written in terms of \(Q_F\) if a characteristic time \(\tau_F\) is introduced. The equation for current (in analogy to Equation 7.3.2) is

\[
I_C = \frac{Q_F}{\tau_F}
\]  

(7.4.1)

Note that all that is formally required to make Equation 7.4.1 accurate is that \(I_C\) must be linearly related to \(Q_F\) (with \(\tau_F\) taken to be constant).

Because \(Q_F\) represents the sum of the magnitudes of the excess minority charge in the quasi-neutral emitter and base regions, its voltage dependence is generally that of a diode, and we may write

\[
Q_F = Q_{F0} \left[ \exp \left( \frac{qV_{BE}}{kT} \right) - 1 \right]
\]

(7.4.2)

where \(Q_{F0}\) is a function of the dopant profiles and device geometry. Since \(Q_F\) is at least somewhat greater than \(Q_{nb}\), \(\tau_F\) must be greater than the base transit time \(\tau_B\) (Equation 7.3.2). Further comments about \(\tau_F\) are deferred until the charge-control model is developed more fully.

The steady-state current flowing in the base lead is proportional to the rate at which \(Q_{nb}\) recombines in the quasi-neutral base plus the rate at which holes are injected into the emitter to replenish \(Q_{pe}\). These two rates are proportional to the diode factor \(\left[ \exp \left( \frac{qV_{BE}}{kT} \right) - 1 \right]\), and therefore (by Equation 7.4.2) proportional to \(Q_F\). It is therefore possible to write a charge-control expression for the input (base) current of the transistor

\[
I_B = \frac{Q_F}{\tau_{BF}}
\]

(7.4.3)

A considerable amount of physical analysis, involving emitter efficiency and the recombination processes for excess carriers in both the emitter and base, would be required to express \(\tau_{BF}\) or \(\tau_F\) in terms of more fundamental parameters. This analysis is not of special value because both \(\tau_{BF}\) and \(\tau_F\) can be obtained in practice from measurements, and also because the charge-control model aims at representing the device for design, not at exploring the physical electronics of transistor models.

Equations 7.4.1 and 7.4.3 can be used to show that the steady-state current gain is simply the ratio of the two characteristic times:

\[
\frac{I_C}{I_B} = \beta_F = \frac{\tau_{BF}}{\tau_F}
\]

(7.4.4)
For example, Equations 7.4.1, 7.4.3 and 7.4.4 can be applied to the prototype transistor of Figure 6.1. If the emitter efficiency is very high in the prototype device, then \( Q_F \approx Q_{nB} \) with \( Q_{nB} = \frac{1}{2} q n(0) x_B A_F \). Since only base recombination is significant for this case, \( \tau_{BF} = \tau_n \) and \( \tau_F = x_B^2 / 2 D_n \) as derived in Equation 7.3.3. Therefore, from Equation 7.4.4 \( \beta_F \) is \( 2 L_n / x_B^2 \) where \( L_n = \sqrt{D_n} \tau_n \). This result for dc current gain can be compared with earlier analysis of the same problem in Section 6.2. There, \( \alpha_F \) was derived as the base transport factor to be \( [1 - (x_B^2 / 2 L_n^2)] \) in Equation 6.2.8. If we use this result for \( \alpha_F \) in the equation \( \beta_F = \alpha_F / (1 - \alpha_F) \), we find an identical value for \( \beta_F \) to that obtained from the charge-control analysis.

A full charge-control model for the bipolar transistor is derived by adding terms to represent the currents that flow because of time variations in stored charge. Clearly, if \( Q_F \) increases with time, there will be a component of base current equal to \( dQ_F / dt \). Likewise, changes in the charges stored at the base-emitter and base-collector junctions \( (Q_{VF} \) and \( Q_{VC}) \) result in added base current. An overall expression for the base current is, therefore,

\[
i_b = \frac{Q_F}{\tau_{BF}} + \frac{dQ_F}{dt} + \frac{dQ_{VF}}{dt} + \frac{dQ_{VC}}{dt} \quad (7.4.5)
\]

The first three components of current in Equation 7.4.5 flow from the base to the emitter, the last flows from the base to the collector. Combining Equation 7.4.1 and 7.4.5 and using Kirchhoff's current law, one obtains a set of charge-control equations for the transistor under active bias:

\[
i_C = \frac{Q_F}{\tau_F} - \frac{dQ_{VC}}{dt} - \frac{dQ_{VF}}{dt}
\]

\[
i_b = \frac{Q_F}{\tau_{BF}} + \frac{dQ_F}{dt} + \frac{dQ_{VF}}{dt} + \frac{dQ_{VC}}{dt} \quad (7.4.6)
\]

\[
i_E = -Q_F \left( \frac{1}{\tau_F} + \frac{1}{\tau_{BF}} \right) - \frac{dQ_F}{dt} - \frac{dQ_{VF}}{dt} \]

We have thus derived a set of linear equations relating currents and charges in a bipolar transistor. These equations are in contrast to the nonlinear expressions that relate currents and voltages in the transistor.

The circuit diagram of Figure 7.15 represents the various terms in Equation 7.4.6. The diode from the base to the emitter passes the steady-state current and has a saturation current \( I_{FS} = Q_{F0} (1 / \tau_F) + (1 / \tau_{BF}) \) as indicated by Equation 7.4.2. The elements storing the charges \( Q_F, Q_{VF}, \) and \( Q_{VC} \) are shown as capacitors with a line across them to indicate that they store charge (like capacitors), but are voltage dependent. Now that a basic set of charge-control equations has been written for a BJT, we are better able to gain a perspective on the limitations of this viewpoint.

The basic premise underlying charge-control analysis is the existence of a constant proportionality between quantity of charge and current. Another way of stating this is that the characteristic times in the charge-control equations must not themselves be functions of characteristic times may be functions of dynamic conditions. Since the current is a function of the charge carriers must disperse through the base-emitter junction and not have the same ratio.

If a dynamic problem for charge is constrained by a base current, then a fair representation of significant error in the characteristic times, that is, the overall error is appreciably greater than the error in the characteristic times, is perfectly adequate. A careful analysis follows.

**Applications of the Charge-Control Model**

Before discussing the biasing of a bipolar transistor, we illustrate the use of Equation 7.4.6 for this purpose is shown in Figure 7.11. The npn transistor under active bias, with hand calculations, several approximations can be considered negligible. The npn transistor under forward bias is the same. Because we have chosen the design conditions, it is also possible to treat
not themselves be functions of charge or of bias voltages. Under this premise, the characteristic times may be derived for dc conditions and then be used to express dynamic conditions. Strictly this premise is not correct; the characteristic times are functions of the charge. For example, to affect current at the collector, minority carriers must disperse through the base region after being injected at the edge of the base-emitter junction. Thus, during transient conditions, collector current does not have the same ratio to base charge as it does in the steady state.

If a dynamic problem is analyzed with the charge-control model, the solutions for charge are constrained to be a time sequence of differing ‘‘steady-state” solutions. Hence, these solutions are sometimes called quasi-static approximations. For the greater part of the transient, however, the charge-control solution is usually a fair representation of the more exact result. The time scale over which there is significant error in the charge-control solution is of the order of the transit time in the base, that is, the order of \( \tau_F \). For most applications the time scale of interest is appreciably greater than \( \tau_F \), and solution by means of the charge-control model is perfectly adequate. A specific example in the next section will make this more clear.

Applications of the Charge-Control Model

Before discussing the bipolar charge-control model further, it is worthwhile to illustrate the use of Equations 7.4.6 in an application. A simple circuit appropriate for this purpose is shown in Figure 7.16a, and the circuit plus charge-control model is sketched in Figure 7.16b. What is sought is the collector-current response of an \( npn \) transistor under active bias when driven by a current source at the base. For hand calculations, several simplifications are appropriate. First, the current \( dQ_{VC}/dt \) can be considered negligible since the base-emitter voltage varies only slightly under forward bias. This simplification is usually made for active-bias operation. Because we have chosen a simple circuit in which the collector voltage is constant, it is also possible to treat the current \( dQ_{VE}/dt \) as negligible.
With these approximations the equation for the base current has only one unknown, the controlled charge \( Q_F \), and takes the form

\[
  i_B = \frac{Q_F}{\tau_{BF}} + \frac{dQ_F}{dt} \quad (7.4.7)
\]

Since \( i_B(t) \) is specified by

\[
  i_B = i_{B1} (t < 0) = i_{B2} (t > 0)
\]

the solution for \( Q_F \) will contain terms associated with both the homogeneous and particular forms of Equation 7.4.7. When the boundary values \( Q_F(t = 0) = i_{B1} \tau_{BF} \) and \( Q_F(t \to \infty) = i_{B2} \tau_{BF} \) are used, the solution becomes

\[
  Q_F = \tau_{BF}[i_{B2} + (i_{B1} - i_{B2}) \exp(-t/\tau_{BF})] \quad (7.4.8)
\]

Under the approximations that we have made, the collector current is given by \( Q_F/\tau_F \) so that its time dependence becomes (using Equation 7.4.4)

\[
  i_C = \beta_F \left[ i_{B2} + (i_{B1} - i_{B2}) \exp(-t/\tau_{BF}) \right] \quad (7.4.9)
\]

The collector current thus changes from its initial to its final value following an exponential function with a characteristic time constant equal to \( \tau_{BF} \) (Figure 7.17).

As mentioned in the previous section, the transient solution given in Equation 7.4.9 is in error for small values of \( t \). At zero time, for example, Equation 7.4.9 predicts an abrupt change in the slope of the collector current equal to \( (i_{B2} - i_{B1})/\tau_F \) whereas the collector current rises more smoothly as calculated from the transient solution.

The foregoing approximations that have been implicit in the analysis of the base transit time and the collector current equations are not entirely unrealistic and are not likely to have any significant adverse effects on the model for times larger than several base transit times.

One such complication that was not considered in previously was the effect of the load resistor \( R_L \) on the collector voltage. The collector voltage is then given by \( V_{CC} - i_C R_L \) as shown in Figure 7.17.
7.4 Charge-Control Model

Figure 7.17  Time variation of collector current in the circuit of Figure 7.16 as calculated from the charge-control model.

\[
\frac{(i_{b2} - i_{b1})}{\tau_f} \quad \text{whereas collector current will not change until the extra electrons injected at the emitter side of the base reach the collector. A more complete analysis that takes account of the distributed nature of base charging predicts that } i_c \text{ does not change at all for times close to } t = 0; \text{ initially, the increments in the base and emitter currents are equal. The collector current first begins to change when } t \text{ reaches the base transit time } \tau_B. \text{ The behavior of } i_c \text{ then rapidly approaches the transient solution predicted by the charge-control model and essentially matches that result for times larger than } t = \tau_f. \text{ (See inset in Figure 7.17.)}
\]

The foregoing example may seem artificial because of the number of simplifications that have been imposed, first in the choice of circuit and second in the approximations that have been made. These physical simplifications, however, have avoided mathematical complications that might tend to obscure the use of the model.

One such complication would arise if, for example, the collector in Figure 7.16a were not connected to an ac ground, but rather was connected to the source through a load resistor \( R_L \). Since \( V_C \) would be variable in this case, the current required to charge \( Q_{VC} \) would not be negligible. The solution for \( i_c \) can be obtained readily if we define an effective capacitance \( C_{fe} \) equal to the average of \( dQ_{VC}/dV_C \) over the collector voltage interval. The solution for \( i_c \) is then found to be equal to Equation 7.4.9 provided that the time constant \( \tau_{bf} \) in Equation 7.4.9 is replaced by [Problem 7.19]:

\[
\tau_{bf}' = \tau_{bf} \left( 1 + \frac{R_L C_{fe}}{\tau_f} \right) \quad (7.4.10)
\]

To derive Equation 7.4.10, it is necessary to note that the presence of the collector capacitance affects the base current much more than it does the collector current.
is also induced. For example, under very high forward bias, the number of holes can rise above the value of the acceptor doping in the base. This condition is termed high-level injection. Equation 8 demonstrates that there is a drop in collector current. If the injected hole concentration is very much greater than the background acceptor density, then quasineutrality will be established through the relation \( p = n = n_t \exp(qV_{BE}/2kT) \). This leads to an expression for \( J_c \) with voltage dependence \( \exp(qV_{BE}/2kT) \).

If the reverse-bias voltage \( V_{CB} \) is increased, then the depletion region in the base will grow, and the integrated hole density will decrease. Through Eq. 8, this must lead to an increase in collector current. The change is found, for the case of uniformly doped base, to be

\[
\frac{dJ_c}{dV_{CB}} = J_c C_{BC}/p_b w
\]

where \( C_{BC} \) is the base–collector depletion capacitance per unit area. The linear expression for \( J_c \) as \( V_{CB} \) is varied, is given by

\[
J_c(V_{CB})/J_c(0) = 1/(1 - V_{CB}/V_A) \approx 1 + (V_{CB}/V_A)
\]

where \( V_A = C_{BC}/(p_b w) \) is known as the Early voltage. The associated base conductance degrades transistor voltage gain, and must be minimized to maintain a suitably large value of \( p_b w \).

The above expressions for \( J_c \) correspond to cases in which electron flows dominate by base transport. This covers most situations encountered in most devices. As described below, additional contributions must be added to the heterojunction bipolar transistors where potential barriers at current flow, or produce injected electron distributions far from equilibrium. In the most aggressively scaled devices, further corrections are required, inasmuch as Eq. 7 is not valid if carriers experience few scattering events as they traverse the base.

**Current.** The hole current which must be supplied from the base is a number of different components, corresponding to recombinations of holes with electrons in different regions of the device. As pictured in Fig. 4, there is recombination in the quasineutral base, in the emitter–base charge region, at the emitter periphery, in the emitter body, and at emitter surface. These key components of the base current are evaluated following.

**Recombination.** Recombination of electrons and holes within the base proceeds by direct photon-induced recombination, via deep levels, or by the Auger effect. Under most circumstances, a recombination rate \( \tau_{rec} \) can be determined such that the net number of recombinations}

**Fig. 4** Schematic components of emitter and base current in \( n-p-n \) transistors.

per unit volume in the base is \( U = (n - n_{eq})/\tau_{rec} \), where \( n \) is the density of electrons injected into the base, and \( n_{eq} \) represents the thermal equilibrium minority carrier concentration. The associated base current is the integral of the recombination density over the base:

\[
J_{b1} = qN_s/\tau_{rec}
\]

where \( N_s \) is the integrated density of excess electrons injected into the base. For the simple transistor with uniform base, \( N_s \) can be directly evaluated as

\[
N_s = [n(0) + n(w) - 2n_{eq}]w/2 = n_t[\exp(qV_{BE}/kT) + \exp(qV_{BC}/kT) - 2]w/2p_b
\]

In normal operation \( V_{BE} \gg kT/q, V_{BC} \ll -kT/q \), only the term involving \( V_{BE} \) is important. \( N_s \) can be easily related to the collector current \( J_c \) for a transistor with uniform doping and composition from Eq. 1 as

\[
N_s = J_c w^2/(2D_nq) = J_c \tau_b/q
\]

Here \( \tau_b \) is the base transit time for electrons. Thus the base current \( J_{b1} \) and the associated current gain \( \beta_b \) limited by this mechanism alone are

\[
J_{b1} = (\tau_b/\tau_{rec})J_c
\]

\[
\beta_b = J_c/J_{b1} = \tau_{rec}/\tau_b
\]

The ratio of base transit time \( \tau_b \) to recombination time \( \tau_{rec} \) can be viewed as the probability of recombination in the base of a given injected electron. In more complex structures, \( N_s \) can also be found to be proportional to \( J_c \), and Eqs. 14–15 continue to hold with an appropriately modified expression for \( \tau_b \). In most modern devices, \( \tau_b \) is of order of picoseconds, whereas \( \tau_{rec} \)
is of order nanoseconds or microseconds; thus $J_{b1}$ is generally quite small.

*Emitter–Base Depletion Region Recombination.* According to the statistics for recombination via deep levels (quantified by Shockley–Read–Hall and others\(^1\)), the effective recombination lifetime for carriers varies strongly with carrier density. The recombination rate reaches a strong maximum under the condition $p = n$, inasmuch as both carriers are relatively plentiful for capture. Within the bipolar transistor, this occurs over a relatively thin region within the emitter–base depletion region, for which $n = p = n_i \exp(qV_{BE}/2kT)$. Integrating the recombination density over the depletion region leads to the net base-current contribution:

$$J_{b2} = q \int U(x)dx = (q\varepsilon /\tau_{ef}) (2\pi kT/q\varepsilon_p) \exp(qV_{BE}/kT)$$

where $\varepsilon_p$ is the electric field at the plane of maximum recombination. This contribution to base current thus has a voltage dependence of $\exp(qV_{BE}/2kT)$, that is, an ideality factor of 2. Correspondingly, the current gain $\beta_2$ limited by this mechanism increases with increasing collector current:

$$\beta_2 = J_c/J_{b2} \sim \exp(qV_{BE}/2kT) \times J_c^{1/2}$$

*Reverse Injection into Emitter.* When the base-emitter junction is forward-biased, holes flow by diffusion (and potentially drift) into the emitter, recombining with electrons in the emitter body and at the emitter surface. The carrier concentration profile differs significantly among devices, according to the thickness of the emitter relative to the hole diffusion length, and according to the recombination velocity at the surface of the emitter. The carrier profile shown in Fig. 3 corresponds to a situation with a thick emitter, where the diffusion length of holes governs their profile. More frequently, bipolar transistors have thin emitters with metallic contacts (causing high surface recombination velocity), for which the density of holes varies approximately linearly with distance within the emitter. In either case, the hole current density can be calculated taking into account diffusion, drift associated with built-in fields, and varying bandgap, with a method directly analogous to the treatment of electron flow (Eqs. 4–8). This results in hole current $J_{b3}$ given by

$$J_{b3} = qD_p \exp(qV_{BE}/kT) \int (n/n_i^2)dx$$

The DC current gain of the transistor, $\beta_3$, as limited by this mechanism, is

$$\beta_3 = J_c/J_{b3} = (D_n/D_p) \int (n/n_i^2)dx / \int (p/n_{ib}^2)dx$$
where $n_{ie}$ and $n_{ib}$ are the intrinsic carrier concentrations in emitter and base, respectively. For uniform doping of the emitter and base, and a thin emitter with infinite surface recombination velocity, this expression yields

$$
\beta_3 = (D_e/D_p)(n_{ie}w_e^{1/2}/p_{ib}w_e^{1/2}) \exp(E_{ge}/kT)$$

(19)

This equation reveals a number of important factors for transistor design. Current gain is directly proportional to the ratio $n_{ie}/p_{ib}$ (emitter to base doping concentration). To maintain suitably high values of current gain, this ratio is typically chosen to be on the order of 1000. For a silicon homojunction transistor, emitter doping levels of $10^{20}/cm^3$ and base doping levels on the order of $10^{17}-10^{18}/cm^3$ are typically used. Also seen in Eq. 19 is the fact that the difference in bandgap energy between emitter and base has a major impact on current gain. In homojunction transistors, the principal bandgap difference stems from the bandgap shrinkage resulting from heavy doping on the emitter side of the device. For high doping levels, the effective bandgap energy is reduced because of the effects on carrier energy of the potentials of the donor or acceptor atoms and their associated fluctuations (bandtailing effect), and because of the binding energy of the associated electron and hole gases. The extent of bandgap shrinkage in silicon has been extensively studied theoretically and experimentally. As a result of this reduction in emitter bandgap, the current gain is reduced according to the factor \(\exp(-\Delta E_{oe}/kT)\), where \(\Delta E_{oe}\) is the bandgap narrowing at the emitter–base junction edge.

The overall base current is the sum of these components, plus, potentially, the additional currents associated with tunneling at the emitter–base junction and emitter–edge currents. Base current components are often difficult to distinguish in specific circumstances, although in general they differ in their dependence on $V_{BE}$, on temperature, on base thickness, and on the ratio of emitter periphery to emitter area.

1.2.2 Charge Storage

The AC and transient characteristics of bipolar transistors are dominated by the charge stored within the device that must be increased or decreased when the bias conditions are changed. Within the charge control model, the base current consists of a DC component that is associated with the instantaneous bias voltages at the device terminals and a transient value given by the derivative of the charge stored in the device. For normal device operation, this corresponds to

$$
I_B(t) = I_{B0}[V_{H1}(t), V_{H2}(t)] + (dQ_B/dt)
$$

$$
I_C(t) = I_{C0}[V_{H1}(t), V_{H2}(t)]
$$

$$
I_I(t) = I_{I0}(t) + I_{i0}(t)
$$

(20)
excess hole and charge stored in the transistor in forward bias.

Moreover, in the charge control approximation, the value of \( Q_B \) is taken to be the value that would apply under steady-state conditions, given the instantaneous values of the terminal voltages. In the spirit of this approximation, from knowledge of the charge distribution in the transistor under steady-state conditions and Eq. 20, the transient and AC performance of the transistor can be calculated.

The charge stored in the transistor, \( \Delta Q_B \), is in excess of the charge contained in the device at zero bias, \( Q_{B0} \), is pictured in Fig. 5. Here the electron and hole distributions are shown for conditions of zero bias and a representative active bias. The excess charge per unit area \( \Delta Q_B \) can be computed by integrating over the entire device the contributions corresponding to excess electron charge density, or the contributions corresponding to excess hole charge. The two resulting values are equal in magnitude, since the overall transistor is neutral.

**Charge Contributions.** Distinct contributions to the charge, \( Q_j \), may be associated with the different regions of the device (termed \( Q_{bc1} \), \( Q_{bc2} \), \( Q_B \), \( Q_{be1} \), and \( Q_{be2} \) in Fig. 5). Each of these contributions is approximately proportional to collector-current density \( J_c \). It is of interest to compute the ratio \( Q_j/J_c \), a quantity with the units of time, which corresponds loosely to the transit time of carriers through the different regions of the device. These are approximately evaluated in the following.

**Emitter Region, \( Q_{be1} \).** Under forward-bias of the base-emitter junction,
excess holes are injected into the emitter with a distribution that depends on the details of the emitter thickness and surface recombination (as described above). For the simple case of thin emitter with metal contact, the charge stored is

\[
Q_{bc1} = \int (p - p_{eq}) \, dx = q w_c n_{eq}^2 \exp(qV_{BE}/kT)/(2n_c)
\]

\[
\tau_{bc1} = Q_{bc1} / J_c = w_c w_p n_{eq}^2 (2D_n n_c n_{th})
\]

\[
(21)
\]

\( \tau_{bc1} \) is a significant contribution to the overall delay of silicon bipolar transistors. To minimize it, thin emitters with high doping are desired, while lightly doped, thin bases are also important (as also required to establish high DC current gain). Another effective strategy is to use a wide bandgap emitter, as described later for heterojunction bipolar transistors (HBTs).

**Emitter-Base Depletion Region, \( Q_{bc2} \).** Charges must be stored at the edges of the depletion region to support the electrostatic fields as the junction voltage is changed. For a change in voltage \( dV_{BE} \), the corresponding change in stored charge is \( C_{BE} dV_{BE} \). \( C_{BE} \) may be approximately evaluated as the capacitance of a \( p-n \) depletion region. The delay time \( \tau_{bc2} \) associated with this charge is

\[
\tau_{bc2} = dQ_{bc2}/dJ_c = C_{BE} (dJ_c/dV_{BE}) = C_{BE} s_m = C_{BE} qJ_c/kT
\]

\[
(22)
\]

\( \tau_{bc1} \) and \( \tau_{bc2} \) are frequently lumped together in a contribution \( \tau_c \) known as the emitter delay.

**Base Region, \( Q_b \).** Charge associated with the electrons injected into the base is neutralized with additional holes added to the base. The overall amount of charge, and the associated delay time, have already been discussed for the calculation of base current. For a uniform base, they are given by

\[
Q_b = \int q(n - n_{eq}) \, dx = J_c w^2/(2D_n)
\]

\[
\tau_b = w^2/(2D_n)
\]

\[
(23)
\]

In modern devices, there are a variety of corrections that must be considered. When significant electrostatic fields are present in the base, electron flow proceeds by drift as well as by diffusion, and \( \tau_b \) is customarily represented as \( \tau_b = w^2/\eta D_n \), where \( \eta \) is an adjustment factor that depends on the magnitude of the electric field present (with \( \eta \approx 2[1 + (q' w/2kT)^{3/2}] \)).

The velocity with which electrons may exit from the base is limited to a value of the order of the saturation velocity, \( v_{sat} \). In addition, in thin bases, diffusive flow is not governed by the simple Fick's law expression; rather, it is limited to a velocity known as effusion velocity or thermionic-emission velocity. This velocity corresponds to the situation in which an entire
thermal electron population is directed from emitter to collector, and there is no backscattered or returning carrier flow. Under such circumstances, an approximate expression for base transit time is given by

\[ \tau_n = \frac{w^2}{(2D_n)} + \frac{w}{v_m}, \]

where \( v_m \) is the velocity at which the electrons exit the base at the collector edge. \( v_m \) is typically given by the diffusion or thermionic-emission velocity of electrons. \( v_m = \left( kT/2 \pi m^* \right)^{1/2} \).

**Collector Region, \( Q_{bc1} + Q_{bc2} \).** As bias conditions are changed, the charge stored in the base-collector depletion region must change, through two mechanisms. In the first mechanism, if \( V_{bc} \) changes, then a depletion charge of \( Q_{bc1} = C_{bc} dV_{bc} \) must be added (or removed) to the base and to the collector edges of the depletion region. Correspondingly, there is delay time \( \tau_{bc1} \) defined as \( dQ_{bc1}/dI_c \). By convention, the charge is computed with the collector terminal incrementally short-circuited to the emitter terminal. As a result, by taking into account the external circuit, which may have series resistances, as \( V_{BE} \) is changed, the variation of \( V_{BC} \) is given by

\[ \Delta V_{BC} = \Delta V_{BE} + I_c (R_E + R_C) \]  

(24)

where \( R_E \) and \( R_C \) are extrinsic parasitic resistances associated with these terminals. The collector current \( I_c \) is given by \( J_c \times \text{area of the emitter}, A_E \). The delay time is thus

\[ \tau_{bc1} = C_{bc} (dV_{bc}/dI_c + R_E A_E + R_C A_E) \]

\[ = C_{bc} (qI_c/kT + R_E A_E + R_C A_E) \]  

(25)

A second mechanism causing a variation of the charge in the base (at constant \( V_{bc} \)) is associated with the finite velocity of electrons within the depletion region. As \( J_c \) increases, the electron density within the collector also increases, and the associated electron charge modifies the space-charge density distributed throughout the collector, from \( N_{1D} \) to \( N_{1D} - J_c/qv_s \) (where \( N_{1D} \) is the donor concentration in the collector region, and \( v_s \) is the electron velocity, typically at its saturated value). The injected electrons act as acceptor dopants in the depletion region. The resultant change in "doping" changes the amount of charge at the base edge of the depletion region. Within the simple approximation for a one-sided p-n junction, the charge \( Q_{bc} \) in the depletion region, and the associated time constant \( \tau_{bc2} \) are

\[ Q_{bc} = \frac{(2qN_{1D} + (V_{CB} + V_{bi})}{(2N_{1D} - J_c/qv_s)} \]  

\[ \tau_{bc2} = dQ_{bc}/dI_c = \left( qI_c/2N_{1D} \right) = w_s/2v_s \]  

(26)

where \( w_s \) is the thickness of the collector depletion region. The result for \( \tau_{bc2} \) corresponds to one-half the time expected for an electron to traverse the collector depletion region, traveling at the saturated drift velocity. The factor of two accounts for the fact that the charge associated with the electrons distributed through the depletion region is doubled, as in the situations in the base and the collector, discussed above. In practice, the base-collector delay is

**Modes of \( V_{bc} \) Change.** Depending on the bias conditions, the base-collector junction can be biased in the base-collector depletion mode, and base-emitter, and base- collector, and base-emitter-emitter mode. The ratios employed to determine current gain \( h_{fe} \) in the base-emitter mode, and base-emitter, and base-emitter-emitter mode, under normal operating conditions, are stored within the transistor. During testing of the device, the ratios are then determined and stored in the database, along with the other information about the device. The overall testing process is then completed.
distributed through the depletion region is terminated partly at the base depletion region edge, and partly at the collector depletion region edge, under conditions of constant voltage drop across the depletion region. For situations in which the velocity of the electrons varies spatially within the collector, the above simple picture may be modified to give

$$\tau_{\text{bc}} = \int \left(1 - \frac{v}{v_c}\right) \psi(v) \, dv$$

(27)

The overall charge stored in the base is the sum of the contributions defined above. In similar fashion, the overall delay, termed $\tau_{\text{ec}}$, the emitter-to-collector delay, is the sum of the contributions described:

$$Q_B = Q_{\text{be}1} + Q_{\text{be}2} + Q_b + Q_{\text{bc}1} + Q_{\text{bc}2}$$

$$\tau_{\text{ec}} = \tau_{\text{be}1} + \tau_{\text{be}2} + \tau_b + \tau_{\text{bc}1} + \tau_{\text{bc}2}$$

(28)

**Modes of Transistor Operation.** The preceding discussion centered on bias conditions in which the emitter-base junction is forward-biased and the base-collector junction is reverse-biased or only moderately forward-biased (called the forward active mode of operation). If the base-emitter junction and base-collector junctions are both reverse-biased, the transistor is in cutoff mode, and no current flows. With the base-collector forward-biased and the base-emitter reverse biased, inverse operation of the transistor is obtained, and emitter current will flow. Under most circumstances, the doping and area ratios employed in the transistor fabrication will lead to a very low value of current gain in this mode of operation, potentially below unity. Finally, if the base-emitter junction and base-collector junctions are both forward-biased, the transistor is said to be in saturation. Representative minority-carrier concentrations in the saturation mode are shown in Fig. 6. The charge stored within the base is considerably greater than what would be stored under normal operation. Also, large amounts of minority-carrier charge are stored in the collector region. The excess charges (holes) must be disposed of during switching operations either by recombination or by extracting them from the base terminal. Typically, they slow the transistor performance dramatically. In saturation, the charge stored is partitioned into contributions ($Q_F$ and $Q_R$) that are associated with emitter and collector. The base charge of these two portions can be determined from the graphical analysis of Fig. 6. The overall form of the charge control equations, including saturation, then becomes:

$$I_B(t) = I_{\text{be}0} |V_{\text{be}}(t)| \, V_B(t) + (dQ_F/dt) + dQ_R/dt$$

(29a)

$$I_C(t) = I_{\text{col}} |V_{\text{bc}}(t)| \, V_C(t) + (dQ_R/dt)$$

(29b)

$$I_E(t) = I_{\text{col}} |V_{\text{be}}(t)| \, V_E(t) + (dQ_F/dt)$$

(29c)
1.2.3 Figures-of-Merit for Transistor Performance

Current Gain Cutoff Frequency \(f_T\). \(f_T\) is the frequency at which the magnitude of the transistor incremental short-circuit current gain, \(h_{fe}\), drops to unity. It is a key estimator of transistor high-speed performance. If we consider AC transistor behavior, assuming small-signal excitation voltages and currents with time dependence \(\exp(j\omega t)\), then within the charge control framework, the incremental current gain \(h_{fe}\) can be calculated as follows:

\[
\begin{align*}
    i_b &= i_{di} + j\omega q_i, \\
    i_c &= i_{di} \\
    |h_{fe}| &= |i_c/i_b| = |i_{di}/(i_{di} + j\omega q_i)| = 1/[(i_{di}/f_{di}) + (j\omega q_i/f_{di})] = 1/[(1/\beta + j\omega t_{re})]
\end{align*}
\]  

(30)

Here we have introduced notation \(i_b\), \(i_c\), etc. to denote small-signal values of the quantities \(I_B\), \(I_C\), etc., and have noted \(q_i/i_{di} = dQ_B/dI_C = \tau_{re}\). The current gain \(h_{fe}\) has the frequency dependence noted in Fig. 7, reaching a low-frequency value of \(\beta\), and dropping at a rate of 6 dB/octave (1/f dependence) at frequencies above the beta-rolloff frequency \(1/(2\pi t_{re})\). The current gain magnitude drops to unity at a frequency \(f_T = 1/(2\pi t_{re})\), an expression arrived at by neglecting \(1/\beta\) compared with unity. The delays studied in Section 1.2.2 thus have a key role in governing the AC current gain of the device. A brief summary of the contributions to \(f_T\) is frequently written, for normal operation at moderate current densities, as

\[
1/(2\pi f_T) = \tau_b + \tau_e + (w_s/2v_s) + (R_E + R_C)C_{BC} + [(C_{BE} + C_{BC})/kT(\eta I_C)]
\]

(31)

The first three contributions are sometimes referred to as \(T_F\) in circuit level models, as described in the following. The last component depends on collector current gain, \(h_{fe}\), etc.

Fig. 7 Reaches vc described.

Fig. 8; the maximum useful to maximum
collector current as $1/I_C$, and leads to a dramatic slowdown of the transistors at low current levels. To distinguish the contributions, it is customary to plot experimental data of $f_1$ vs $I_C$ in the form of $1/(2\pi f_1)$ vs $1/I_C$, as shown in Fig. 8; the slope can be identified with $kT/\alpha (C_{\text{Hr}} + C_{\text{Hc}})$. As the current $I_C$ reaches very high levels, $f_1$ drops ($\tau_{\text{cc}}$ rises) due to the base pushout effect described below.

Maximum Frequency of Oscillation ($f_{\text{max}}$). This is the frequency at which the maximum available power gain of the transistor drops to unity. $f_{\text{max}}$ is widely useful to estimate power gain, since over a wide range of frequencies, maximum available power gain, $G_p$, follows the relation

$$G_p = (f_{\text{max}} / f)^2$$

(32)

$f_{\text{max}}$ is different from (and typically larger than) $f_1$, because in addition to current gain, $f_{\text{max}}$ takes into account the possibility of voltage gain. A simple analysis based on the simplified hybrid pi equivalent circuit for the bipolar transistor\(^1\) provides the basis for estimating the factors important to $f_{\text{max}}$. The
To problem is to find \( Q_{BC} \), the variation in the base charge (associated with the base-collector depletion region).

Approach: Find the induced charge because of the sheet \( \Delta n(x) \) in the width \( \Delta x \) and integrate from \( x = 0 \) to \( x = W \).
The voltage across the B-C junction is \((V_{CB} - V_{bi})\) is constant (not perturbed by the change)

\[
\text{Caused by the charge } \Delta E^+ \]

But the charge perturbs the electric field such that

\[
\Delta E^+ - \Delta E^- = \frac{q \Delta n(x) A x}{\varepsilon} \quad \text{(Gauss's Law)}
\]

Also \(\Delta E^+ \cdot x + \Delta E^- (w-x) = 0 \quad \text{(no change in voltage)}\)

\[
\therefore \Delta E^+ = -\frac{\Delta E^+ \cdot x}{w-x}
\]

\[
\therefore \Delta E^+ - (\Delta E^+ \cdot x) = \frac{q \Delta n(x) A x}{\varepsilon}
\]

But \(\Delta n(x) = \frac{Jc}{q W(x)}\)

\[
\therefore \Delta E^+ \left(1 + \frac{x}{w_c-x}\right) = \frac{q^2 Jc \cdot \Delta x}{q W(x) \varepsilon}
\]
\[ \Delta e^+ = \Delta Q_{bc2} = \int J_c \left( \frac{w_c - x}{w_c} \right) \frac{dx}{v(x)} \]

\[ \Delta Q_{bc2} = J_c \left( 1 - \frac{x}{w_c} \right) \frac{1}{v(x)} \, dx \]

\[ Q_{bc2} = J_c \int_0^{w_c} \left( 1 - \frac{x}{w_c} \right) \frac{1}{v(x)} \, dx \]

\[ \frac{dQ_{bc2}}{dJ_c} = \tau_{bc2} = \int_0^{w_c} \left( 1 - \frac{x}{w_c} \right) \frac{1}{v(x)} \, dx \]
1.2.3 Figures-of-Merit for Transistor Performance

Current Gain Cutoff Frequency \( (f_T) \). \( f_T \) is the frequency at which the magnitude of the transistor incremental short-circuit current gain, \( h_{fe} \), drops to unity. It is a key estimator of transistor high-speed performance. If we consider AC transistor behavior, assuming small-signal excitation voltages and currents with time dependence \( \exp(j\omega t) \), then within the charge control framework, the incremental current gain \( h_{fe} \) can be calculated as follows:

\[
\begin{align*}
    i_n &= i_{n0} + j\omega q_b \\
    i_c &= i_{c0}
\end{align*}
\]

\[
|h_{fe}| = |i_c/i_n| = |i_{c0}/(i_{n0} + j\omega q_b)| = 1/|(i_{n0}/i_{c0}) + (j\omega q_b/i_{c0})| = 1/(1/\beta + j\omega \tau_{cc})
\]

Here we have introduced notation \( i_n, i_c \), etc. to denote small-signal values of the quantities \( I_n, I_C \), etc., and have noted \( q_b/i_{c0} = dQ_B/dI_C = \tau_{ce} \). The current gain \( h_{fe} \) has the frequency dependence noted in Fig. 7, reaching a low-frequency value of \( \beta \), and dropping at a rate of 6 dB/octave \( (1/f) \) dependence) at frequencies above the beta-rolloff frequency \( 1/(2\pi\beta\tau_{cc}) \). The current gain magnitude drops to unity at a frequency \( f_T = 1/(2\pi\tau_{cc}) \), an expression arrived at by neglecting \( 1/\beta \) compared with unity. The delays studied in Section 1.2.2 thus have a key role in governing the AC current gain of the device. A brief summary of the contributions to \( f_T \) is frequently written, for normal operation at moderate current densities, as

\[
1/(2\pi f_T) = \tau_b + \tau_c' + (w_c/2\nu_s) + (R_E + R_C)C_{BC} + [(C_{BE} + C_{BC})kT/(qI_C)]
\]

The first three contributions are sometimes referred to as \( T_F \) in circuit level models, as described in the following. The last component depends on collector current as \( 1/I_C \), and at low current levels. To distill experimental data of \( f_T \) vs \( I \), Fig. 8; the slope can be identical at very high levels, \( f_T \) as described below.

Maximum Frequency of Oscillation \( f_{max} \): maximum available power gain useful to estimate power \( g \) maximum available power \( g \)

\( f_{max} \) is different from (and t current gain, \( f_{max} \) takes into analysis based on the simpli transistor provides the basis
Collector current as $1/I_C$, and leads to a dramatic slowdown of the transistors at low current levels. To distinguish the contributions, it is customary to plot experimental data of $f_T$ vs $I_C$ in the form of $1/(2\pi f_T)$ vs $1/I_C$, as shown in Fig. 8; the slope can be identified with $kT/q(C_{BE}+C_{BC})$. As the current $I_C$ reaches very high levels, $f_T$ drops ($\tau_{ec}$ rises) due to the base pushout effect described below.

**Maximum Frequency of Oscillation ($f_{max}$).** This is the frequency at which the maximum available power gain of the transistor drops to unity. $f_{max}$ is widely useful to estimate power gain, since over a wide range of frequencies, maximum available power gain, $G_p$, follows the relation

$$G_p = (f_{max}/f)^2$$

(32)

$f_{max}$ is different from (and typically larger than) $f_T$, because in addition to current gain, $f_{max}$ takes into account the possibility of voltage gain. A simple analysis based on the simplified hybrid pi equivalent circuit for the bipolar transistor provides the basis for estimating the factors important to $f_{max}$. The
1. First consider diodes

A. Long Base diode (has only capacitive delay)

$$C = \frac{dQ}{dV} \quad \text{space charge capacitance} = \frac{e}{w(x)}$$

$$q \phi = \int_{0}^{\infty} \Delta t(x) dx = \frac{q A}{\int_{0}^{\infty}} \phi(x) e^{-x/L_p} dx$$

$$Q = q A \int_{0}^{\infty} \phi(x) dx$$

$$I_p(x=0) = q A \int_{0}^{\infty} \phi(x) \frac{1}{L_p}$$

$$G_s = \frac{I_p(x=0)}{I_p} = \frac{L_p}{L_p}$$

Current in charge control analysis is always considered to be the rate of net hole current have to be supplied to maintain the charge distribution

$$\frac{dQ}{dV} = \frac{dE}{dV} \quad L_p = \frac{L_p}{E} \quad \text{(defines a capacitance, but is not the capacitance seen at the terminals)}$$
Time dependent continuity e.g.

\[ \frac{\partial \rho}{\partial t} = -9A \frac{\partial \rho}{\partial t} - \frac{\partial \rho}{\partial x} \]

\[ \frac{\partial (\rho \phi)}{\partial t} = -\frac{\partial \phi}{\partial x} + D \rho \frac{\partial^2 \rho}{\partial x^2} \]  
(assumes only diffusion)

Assume a solution \( \rho = \alpha \rho(0) e^{\frac{-j\omega t}{2}} + \beta \rho \rho^* e^{j\omega t} \)

\[ j \omega \rho = -\frac{\partial \rho}{\partial x} + D \rho \frac{\partial^2 \rho}{\partial x^2} \]

solution: \( \rho = C_1 e^{-\frac{j \omega}{
\frac{\partial \rho}{\partial x}} + C_2 e^\frac{j \omega}{x} \]

\[ L = \left[ j \omega + \frac{1}{\kappa \rho} \right]^{-\frac{1}{2}} \]  
(frequency dependent diffusion length)

\[ C_2 = 0 \]  
for physical reasons

\[ C_1 = \rho \rho(0) \]

\[ \rho = \rho \rho(0) e^{-j \sqrt{\frac{1}{\kappa \rho} + \frac{1}{\kappa \rho}}} \]

\[ \frac{\partial \rho}{\partial x} = \rho \rho(0) \frac{1}{\sqrt{\kappa \rho \rho(0)}} \]

Small-signal approximation: \( V_c \approx V_{dc} + \frac{V}{\sqrt{\kappa \rho \rho(0)}} \)

\[ \rho \rho(0) = \rho \rho(0) + \rho \rho(0) = \rho \rho(0) \frac{\sqrt{V_{dc} + \frac{V}{\sqrt{\kappa \rho \rho(0)}}}}{\sqrt{V_{dc} + \frac{V}{\sqrt{\kappa \rho \rho(0)}}}} \]

since we are assuming \( \frac{V}{\sqrt{\kappa \rho \rho(0)}} \) to be small, \( \frac{V}{\sqrt{\kappa \rho \rho(0)}} \approx 1 + \frac{\sqrt{V_{dc}}}{\sqrt{\kappa \rho \rho(0)}} \)

\[ \rho \rho(0) = \rho \rho(0) \left( 1 + \frac{\sqrt{V_{dc}}}{\sqrt{\kappa \rho \rho(0)}} \right) \]

\[ \frac{\partial \rho}{\partial x} = \rho \rho(0) \left( \frac{\sqrt{V_{dc}}}{\sqrt{\kappa \rho \rho(0)}} \right) \]

We know \( I_{dc} = \rho \frac{\partial \rho}{\partial x} \rho(0) \Rightarrow \rho \rho(0) = \frac{\sqrt{I_{dc}}}{\sqrt{\frac{\rho \rho(0)}}} \frac{\sqrt{\kappa \rho \rho(0)}}{\sqrt{\kappa \rho \rho(0)}} \]

\[ Y = \frac{\gamma}{\kappa \rho \rho(0)} = \frac{\gamma I_{dc}}{\sqrt{1 + j \omega \rho \rho(0)}} \]

\[ Y = \frac{\gamma}{\kappa \rho \rho(0)} = \frac{\gamma I_{dc}}{\sqrt{1 + j \omega \rho \rho(0)}} \]

\[ Y = \frac{\gamma I_{dc}}{\sqrt{1 + j \omega \rho \rho(0)}} \]

\[ Y = \frac{\gamma I_{dc}}{\sqrt{1 + j \omega \rho \rho(0)}} \]
For sufficiently small frequencies,

\[ Y = \frac{1}{\tau} \left[ 1 + \frac{j\omega \tau}{2} \right] = G + j\omega C \]

\[ C = \frac{\varepsilon_p}{2\tau} \quad \text{This is the diffusion capacitance.} \]

We have ignored the depletion region capacitance.

\[ Y_{\text{tot}} = \frac{1}{\tau} + j\omega (C_{\text{diff}} + C_{\text{dep}}) \]

\[ Y = \frac{1}{\tau} + j\omega \left( \frac{\varepsilon_p}{2\tau} + \frac{\varepsilon_A}{\omega} \right) \]
Apparent Capacitance = \( \frac{dC_0}{dV_0} = \frac{V_0}{R_e} \)

\[ \text{Cdiff} = \frac{1}{2} \frac{\frac{dV}{R_e}}{\frac{dV}{R_e}} \]

Using \( V = G + j \omega X \) allows an equivalent circuit to yet y need to solve time dependent continuity eq.

\[ \frac{dP}{dt} = \frac{1}{2} \nabla \cdot \mathbf{F} + \text{G} \cdot \mathbf{F} \]

\[ P = P_n + P_m, \quad P_m = \beta \rho e^{\omega t} \]

Small signal analysis

\[ 0 = P_m + \frac{\partial}{\partial t} P_{0n} (1 + \frac{V_{0n}}{R_T}) \]

\[ P_{0n} (e^{\frac{V_{0n}}{R_T}}) = P_{0n} (1 + \frac{V_{0n}}{R_T}) \]

\[ Y = \frac{1 + j \omega \tau_p}{R_e} \quad \text{at low frequency} \quad \frac{1 + j \omega \tau_p}{R_e} \frac{2 \tau_e}{2 \tau_e} \]

For high freq, \( Y = \sqrt{\frac{j \omega \tau_p}{R_e}} = \sqrt{\frac{\omega \tau_p}{2 R_e}} + j \sqrt{\frac{\omega \tau_p}{2 R_e}} \]

\[ = \frac{1}{R_e} \sqrt{\frac{\omega \tau_p}{2}} + j \frac{\omega \tau_p}{R_e} \sqrt{\frac{\tau_p}{2 \omega}} \]

Both small signal resistance & capacitance decrease with \( \omega \) as \( \frac{1}{\omega^2} \).
\[ \frac{dn}{dt} = \frac{n_0^2}{A} \quad \text{(Assuming negligible recombination)} \quad A > k = 0 \]

Assume, \( n(x,t) = n_0(x) + n_w(x) e^{jut} \) (Diffusion wave)

Considering only \( n_w \),

\[ \frac{d^2 n_w}{dx^2} = \frac{n_w}{\lambda_e^2} \quad \text{where} \quad \frac{1}{\lambda_e} = \sqrt{\frac{m_e}{2e}} \]

\[ n_w(x) = C e^{-\frac{x^2}{\lambda_e^2}} \]

Apply Shockley B.C. \( n_w(x_0) = 0 \)

\[ n_w(x) = n_w(0) \frac{\sinh \left[ \frac{\omega_0 x}{\lambda_e} \right]}{\sinh \left( \frac{\omega_0 x_0}{\lambda_e} \right)} \]

\[ J_w(0) = g n_0 \frac{dn_w}{dx} \bigg|_{x=0} = -g n_0 \frac{n_w(0)}{\lambda_e} \coth \left( \frac{\omega_0}{\lambda_e} \right) \]

Low Frequency case:

\[ \frac{1}{\lambda_e} = \sqrt{\frac{m_e}{2e}} = 1 + \frac{j \sqrt{\omega}}{2 \alpha_e} \]

\[ \text{if} \quad \omega_0 \ll \alpha_e, \quad \frac{1}{\lambda_0} = \sqrt{\frac{\omega}{2 \alpha_e}} \]
for \( \frac{\Delta p}{p} \ll 1 \), gaining \( \alpha \coth \alpha = 1 + \frac{\alpha^2}{3} \), \( o(x^5) \).

\[
J_w(0) = -\frac{g}{D} \Delta n_w(0) \frac{1}{\Delta w_p} \frac{\Delta p}{x} \coth \frac{\Delta p}{x} \left( \text{injected AC current} \right)
\]

\[
= -\frac{g}{D} \Delta n_w(0) \left[ 1 + j\omega \Delta w_p \right]
\]

Small signal approximation:

\[
\text{transmission gain:} \quad \Gamma_w = V_w / V_{in}
\]

\[
J_w(0) = \left[ G + j\omega \right] V_w
\]

\[
G = \frac{\Delta n_w(0)}{\Delta w_p} \frac{g}{3kT}, \quad C = \frac{\Delta n_w(0) \Delta w_p}{3kT}, \quad \text{just}
\]

But we know that:

\[
\frac{C}{V_w} = \frac{g}{2} \Delta n_w(0) \frac{\Delta w_p}{kT}
\]

\[
\text{Diff:} = \frac{2}{3} \left( \text{apparent} \right)
\]

Collector Current

What is the delay in \( J_c(w) \) relative to \( J_e(w) \)?

\[
J_c(\Delta w_p) = J_w(\Delta w_p)
\]

\[
J_c(\Delta w_p) = \frac{g}{D} \Delta n_w(0) \frac{1}{\sinh \left( \frac{\Delta w_p}{x} \right)}
\]

\[
J_c(\Delta w_p) = \frac{g}{D} \Delta n_w(0) \left[ 1 + \frac{j\omega \Delta w_p}{3kT} \right] \left( \text{expanded sinh} \left( \frac{\Delta w_p}{x} \right) \right)
\]

\( J_{Bw} = \text{Base current as } f(w) \)

\[
J_{ew} - J_w = J_{Bw}
\]

\[
J_{ew} = \frac{g}{D} \Delta n_w(0) \left[ 1 + \frac{j\omega \Delta w_p}{3kT} \right] \left( \text{expanded sinh} \left( \frac{\Delta w_p}{x} \right) \right)
\]
\[ J_{\omega} = -\frac{\delta \in e \mu(\omega)}{\omega \delta p} \left( \frac{j \omega \omega^2}{2 \delta e} \right) \]

\[ J_{\omega} = -j \omega \frac{1}{2} \delta \mu(\omega) \omega^2 \]

\[ I = q A \rho(x) \nu = q A \delta \rho \frac{dx}{dx} \]

\[ \nu = \frac{I}{q A \rho} \]

Total delay: \[ \int_{0}^{\omega_0} \frac{dx}{\text{velocity}} \]

\[ \mathcal{V}_0 = \frac{1}{I} \int_{0}^{\omega_0} \delta \rho \, dx \]

\[ \mathcal{V}_0 = \frac{Q_s}{I} \quad (\text{transit time}) \]
\[ J_{aw} = J_{ew} - J_{cw} = j \omega \frac{1}{2} q n_s(n_e) \omega e \]

\[ J_{aw} = j \omega C_R \]

\[ J_{cw} = e \int G - j \omega C_R \int V_w \]

**AC Input current**:

\[ i_R = i_{aw} = 3 \mu A \]

- **i_R** supplies all stored charge variations in the base
  1) \( i_{bias} \) (stored minority charge)
  2) charging \( C_{be} \)
  3) charging \( C_{bc} \)

\[ \text{Depletion charge contributed by } V_{bc} \]

\[ \text{Depletion charge controlled by } J_c \]
2.4) *High-Frequency Diffusion Currents*

2.4.1) *Diffusion Waves*

At high frequencies, the \( \frac{\partial n}{\partial t} \) and \( \frac{\partial p}{\partial t} \) terms in the continuity equations (2-21) can no longer be set to zero.

In the present Section we continue to neglect recombination/generation effects, and we continue to assume that the minority carrier flow is by pure diffusion, with zero electric field. The continuity equations in one dimension then assume the simple form

\[
\frac{\partial n}{\partial t} = -\frac{\partial}{\partial x} \left( J_n \right) = +D_n \frac{\partial^2 n}{\partial x^2} ,
\]

\[
\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x} \left( J_p \right) = +D_h \frac{\partial^2 p}{\partial x^2} .
\]

(10a)

(10b)

We look here for solutions of the form

\[
n(x,t) = n_0(x) + n_\omega(x)e^{i\omega t} ,
\]

\[
p(x,t) = p_0(x) + p_\omega(x)e^{i\omega t} .
\]

(11a)

(11b)

The dc parts of these are the same as in Sec. 2.2. If we insert the ac parts of (11a) into (10a), the result may be written in the form

\[
\frac{d^2 n_\omega}{dx^2} = \frac{n_\omega}{\lambda_e^2} ,
\]

(12)

where

\[
\frac{1}{\lambda_e} = \sqrt{\frac{\omega}{D_n}} = (1 + i) \sqrt{\frac{\omega}{2D_e}} = \frac{1 + i}{\Lambda_e} ,
\]

where \( \frac{1}{\Lambda_e} = \sqrt{\frac{\omega}{2D_e}} .
\)

(13a,b)

The solutions of (12) are evidently linear superpositions of waves of the two forms

\[
n_\omega(x) = C \exp \left( \pm \frac{x}{\lambda_e} \right) = C \exp \left( \pm \frac{(1+i)x}{\Lambda_e} \right) ,
\]

(14)
which correspond to attenuated waves towards either the left (+) or the right (−). An analogous solution is found for the hole distribution.

The simplest case is that of an infinitely long semiconductor, in which case we have (for electron diffusing into the p-type semiconductor extending to the right)

\[
n_\omega(x) = n_\omega(0) \exp \left( \frac{x}{\Lambda_e} \right) = n_\omega(0) \exp \left( \frac{(1+i)x}{\Lambda_e} \right).
\]  

(15)

To obtain an idea about the overall scale of the attenuation effects, assume, as an example, \(D_e = 100 \text{ cm}^2/\text{s}\), and \(f = \omega/2\pi = 3 \text{ GHz}\). This yields \(\Lambda_e \approx 10^{-4} \text{ cm} = 1 \mu\text{m}\).

We consider here again the case of a semiconductor of finite thickness \(w_{Bp}\), at the end of which the electron concentration is pinned at its equilibrium value. Inasmuch as the equilibrium density is a steady-state density, we have the boundary condition

\[
n_\omega(w_{Bp}) = 0.
\]  

(16)

The appropriate solution of (12) may be written

\[
n_\omega(x) = n_\omega(0) \frac{\sinh \left( \frac{(w_{Bp} - x)/\Lambda_e}{w_{Bp}/\Lambda_e} \right)}{\sinh \left( \frac{w_{Bp}/\Lambda_e}{w_{Bp}/\Lambda_e} \right)},
\]  

(17)

where \(n_\omega(0)\) is the complex amplitude of the ac part of the electron concentration at the entrance into the p-type semiconductor. We leave its value open for now.

### 2.4.2) AC Currents

Associated with an electron distribution of the form (11a) is a current density of the form

\[
J_{ex}(x,t) = J_0(x) + J_\omega(x)e^{i\omega t}.
\]  

(20)

We are interested here only in the ac part, which is given by

\[
J_\omega(x) = qD_e \frac{dn_\omega}{dx}.
\]  

(21)

If we insert (17) and evaluate the derivative at \(x = 0\), we obtain the complex amplitude of the ac part of the diffusion current entering the p-type semiconductor.
\[ J_\omega(0) = -\frac{qD_e n_\omega(0)}{\lambda_e} \coth \left( \frac{w_{\text{BP}}}{\lambda_e} \right) . \]  

(22)

Evidently, the current amplitude is complex, indicating that the current has both a conductive (real) and capacitive (imaginary) part.

We wish to determine here the capacitive part to the lowest order. To this end we assume that the frequency is sufficiently low that \( \Lambda_e >> w_{\text{BP}} \). In that limit we may expand the hyperbolic cotangent according to its Taylor expansion, which may be written

\[ \alpha \cdot \coth \alpha = 1 + \frac{\alpha^2}{3} - \frac{\alpha^4}{45} + O(\alpha^4) . \]  

(23)

If we keep only terms up to \( \alpha^2 = \left( \frac{w_{\text{BP}}}{\Lambda_e} \right)^2 \), (22) becomes

\[ J_\omega(0) = -\frac{qD_e n_\omega(0) w_{\text{BP}}^2}{3D_e} \left[ 1 + i \omega \frac{w_{\text{BP}}^2}{3D_e} + O(\omega^2) \right] . \]  

(24)

The value of \( n_\omega(0) \) obviously depends on the magnitude of the ac voltage. We set

\[ V_f(t) = V_0 + V_\omega e^{i\omega t} , \]  

(25)

and we assume again Shockley boundary conditions, which may be written

\[ n_p(0,t) = n_{p0} \exp \left( \frac{qV_f(t)}{kT} \right) = n_{p0} \exp \left( \frac{qV_0}{kT} \right) \exp \left( \frac{qV_\omega e^{i\omega t}}{kT} \right) \]  

(26)

The second exponential may always be expanded:

\[ \exp \left( \frac{qV_\omega e^{i\omega t}}{kT} \right) = 1 + \frac{qV_\omega}{kT} e^{i\omega t} + \left( \frac{qV_\omega}{kT} \right)^2 e^{2i\omega t} + O(V_\omega^3) . \]  

(27)

Note that successive terms in this expansion correspond to terms at multiples of the fundamental frequency \( \omega \). These terms give rise to additional terms in both the electron concentration and the electron current, at the various multiples of the fundamental frequency. The treatment of those higher terms is the same as that of the fundamental frequency terms, except that \( \omega \) must everywhere be replaced by the appropriate multiple.

We neglect those higher-frequency terms here, and write
\[
\exp\left(\frac{qV_\omega e^{i\omega t}}{kT}\right) = 1 + \frac{qV_\omega}{kT} e^{i\omega t}.
\]

We may then write (24) as

\[
n_p(0,t) = n_0 + n_\omega e^{i\omega t},
\]

where

\[
n_0 = n_{p0} \exp\left(\frac{qV_0}{kT}\right)
\]

and

\[
n_\omega = n_{p0} \exp\left(\frac{qV_0}{kT}\right) \frac{qV_\omega}{kT} = n_0 \frac{qV_\omega}{kT}.
\]

If we insert (31) into (24), we obtain

\[
J_\omega(0) = -\frac{qD_e n_0}{w_{Bp}} \frac{q}{kT} \left[ 1 + i\omega \frac{w_{Bp}^2}{3D_e} \right] V_\omega.
\]

This may always be written in the form

\[
J_\omega(0) = -(G + i\omega C)V_\omega,
\]

where

\[
G = \frac{qD_e n_0}{w_{Bp}} \frac{q}{kT}
\]

is a differential conductance and

\[
C_{\text{diff}} = \frac{q n_0 w_{Bp}}{3} \frac{q}{kT}
\]

is a differential capacitance, the electron part of what is commonly called the diffusion capacitance of the p–n junction. Analogous contributions to the differential conductance arise from the hole diffusion into the n–type side.

The minus sign in (33) arises again from our sign convention of defining electrical current in the +x direction as positive.
This seemingly bizarre behavior is simply a consequence of the fact that any rise in emitter current appears at the collector with a delay of one transit time. Note that the capacitance seen at the emitter terminal is always positive, namely

\[ C_{\text{diff}} = 2C_B / \beta. \]  

**9.3) Equivalent Circuit Representation.**

The overall behavior of the diffusion currents is easily represented by generalizing the equivalent circuit of Lecture #6, containing a current conveyor isolating the emitter from the collector. We must evidently add the base storage capacitance between the emitter and the base, and the complex transconductance in (9-28) with its negative capacitance between the emitter and the current conveyor:

Note that this equivalent circuit describes only the diffusion currents. To this circuit we must add the space charge layer capacitances on both the emitter and the collector side, as well as the contributions from the back injection current and the base recombination currents. These additions are left as an exercise to the reader.
is a differential capacitance, the electron part of what is commonly called the diffusion capacitance of the p-n junction. Analogous contributions to the differential conductance arise from the hole diffusion into the n-type side.

**Collector Currents:**

The above calculation pertained to the ac diffusion current entering the base from the emitter. At finite frequencies, this current is not equal to the current leaving the base at the collector. If we re-calculate the ac diffusion current density at \( x = w \), rather than at \( x = 0 \), we obtain, instead of (9-11),

\[
J_{C_\omega} = -J_{\omega}(w_B) = \frac{qD_en_\omega(0)}{\lambda_e} \frac{1}{\sinh[w_B/\lambda_e]}.
\]  

(9-25)

An expansion by powers of the frequency, yields, instead of (9-13),

\[
J_{C_\omega} = \frac{qD_en_\omega(0)}{w_B p} \left[ 1 - i\omega \frac{w_B^2}{6D_e} + O(\omega^2) \right].
\]  

(9-26)

The difference between the two currents is of course the base current:

\[
J_{B_\omega} = J_{E_\omega} - J_{C_\omega} = \frac{qD_en_\omega(0)}{w_B p} \left[ i\omega \frac{w_B^2}{2D_e} + O(\omega^2) \right] = i\omega \frac{1}{2} qn_\omega(0) w_B.
\]  

(9-27)

We note that the base current is purely capacitive. A little reflection shows that the capacitance is simply that capacitance that is associated with the increment in electron-hole pair storage in the base region, as a result of the increase in \( V_{BE} \) by \( V_\omega \). It is therefore called the (base) storage capacitance,

\[
C_B = \frac{qn_0w_B}{2} \frac{q}{kT}.
\]  

(9-21)

With the help of (9-23) and (9-21), we may re-write the expression (9-26) for the collector current density as

\[
J_{C_\omega} = \left[ G - i\omega \frac{C_B}{3} \right] V_\omega.
\]  

(9-28)

Note that the capacitive part of this corresponds to a negative capacitance!
The central idea is the following. As the voltage is increased, the current will build up until the traveling negative space charge causes the field at the injecting contact (the emitter) to drop to zero.

In those cases that are of practical interest, the fields throughout most of the semiconductor body are large enough that the current is carried by drift. We shall therefore neglect diffusion altogether, even in the low-field region near the injecting electrode, where strictly speaking, diffusion cannot be neglected. It can be shown that this leads to only a small error in the overall current-voltage relation, but it vastly simplifies the mathematical treatment.

Our point of departure is Poisson’s Equation

\[ \frac{d\mathcal{E}}{dx} = \frac{\rho}{\varepsilon} = \frac{J_e}{\varepsilon v_{sat}}. \]  \hspace{1cm} (13-15)

We integrate, with the initial condition \( \mathcal{E}(0) = 0 \), which yields

\[ \mathcal{E}(x) = \left[ \frac{J_e}{\varepsilon v_{sat}} \right] x. \]  \hspace{1cm} (13-16)

Integrating once more yields the voltage:

\[ V = \left[ \frac{J_e}{2\varepsilon v_{sat}} \right] L^2. \]  \hspace{1cm} (13-17)

We solve for \( J_e \) to obtain the current-voltage characteristic

\[ J_e = 2\varepsilon v_{sat} V/L^2, \]  \hspace{1cm} (13-18)

a linear (Ohm’s-Law-type) dependence. The characteristic feature of this relation is the quadratic rather than linear dependence on length.

13.6) Velocity Saturation Limits on Transistor Speed

The Finite Collector Transit Time

The effect of the finite collector transit time is quite different from that of the finite base transit time, because inside the base-to-collector depletion layer the electrons travel in a medium where they are not neutralized by majority carriers. In effect, the electrons are traveling inside a capacitor. While an electron is in transit inside a capacitor, an electrical current is actually flowing in the leads to that capacitor.

The current is most easily calculated by not considering a single electron, but the effect of an entire sheet of charges, at a distance \( x \) from the left-hand plate, moving with the speed \( v \). (See Fig. 6)
We assume that the electron concentration inside the capacitor is a function of \( x \) only, \( n(x) \). A sheet of infinitesimal thickness \( \Delta x \) at the position \( x = X \) then carries the sheet charge density \( \Delta Q = qn(X) \Delta x \). This sheet contributes an electric field \( \Delta E_+ \) in the region \( 0 < x < X \), and a different field \( \Delta E_- \) in the region \( X < x < w \). From Gauss' Law,

\[
\Delta E_+ - \Delta E_- = -qn(X) \Delta X/e. \tag{13-19}
\]

Also, if the voltage between the capacitor plates is held constant by the external voltage source, we have further

\[
X \cdot \Delta E_+ + (w - X) \cdot \Delta E_- = 0. \tag{13-20}
\]

As the space charge distribution travels, the total charge contained in our sheet will of course remain the same, \( d[n(X) \Delta X]/dt = 0 \). By taking the time derivative of (13-19) and (13-20), one then finds readily

\[
d\Delta E_+/dt = d\Delta E_-/dt = v(\Delta E_+ - \Delta E_-)/w = -q \cdot v \cdot n(X) \cdot \Delta X/e. \tag{13-21}
\]

Here \( v = dX/dt \) is the speed with which the sheet travels. Associated with this time dependence of the two \( \Delta E \)'s is a displacement current density

\[
\Delta J = d\mathcal{D}/dt = (1/e) \cdot \varepsilon dE/dt = -q \cdot v \cdot n(X) \cdot \Delta X/w. \tag{13-22}
\]
But this is of course the current per unit capacitor area flowing in the external leads of the capacitor, due to the contribution of our sheet. The total current per unit area is obtained by integration over all sheets:

\[ J = -q\frac{(v/w)}{\int n(X) \, dX} \quad (\star13-23) \]

Attenuation and Phase Shift of a Traveling Electron Wave

We apply (\star13-23) to a traveling electron wave of the form

\[ n(x,t) = n_0 \exp[j\omega(t - x/v)]. \quad (\star13-24) \]

It obviously corresponds to a wave of (angular) frequency \( \omega \) traveling with the uniform speed \( v \). The convection current in the plane \( x = 0 \) is evidently

\[ J_0 = -qn_0v \exp(j\omega t). \quad (\star13-25) \]

This is the current density that would be flowing if the capacitor were infinitesimally thin and the transit time of the electrons through the capacitor were zero. With the help of (\star13-25) we may write (\star13-24) as

\[ n(x,t) = -(J_0/v) \exp(-j\omega x/v). \quad (\star13-26) \]

If this is inserted into (\star13-23), and the integration executed, one finds readily

\[ J = J_0 \left[ \exp(-j\omega t) - 1 \right] / (-j\omega t), \quad (\star13-27) \]

where we have introduced the electron transit time through the capacitor,

\[ \tau = w/v. \quad (\star13-28) \]

The expression (\star13-27) is easily transformed into the product of an amplitude factor and a phase factor:

\[ J = J_0 \left[ \sin(\omega\tau/2)/(\omega\tau/2) \right] \exp(-j\omega\tau/2). \quad (\star13-29) \]

The two factors following \( J_0 \) indicate the attenuation and the phase shift of the current entering the capacitor by the finite transit time through the capacitor.

We note first of all that the signal delay is only one-half the transit time of the electrons themselves. We also note that there is an attenuation, due to destructive interference between different portions of the travelling wave. For \( \omega = 2\pi/\tau \) the amplitude factor is zero, and no current is collected at all. This is the case when the
wavelength $\lambda = 2\pi v/\omega$ of the traveling wave is equal to the capacitor plate separation $w$.

Clearly, a transistor is restricted to frequencies significantly below $\omega = 2\pi / \tau$ for useful operation. However, we shall see later that one device, the Impact Avalanche Transit Time (IMPATT) diode actually utilizes the strong phase shift just below that frequency.

For sufficiently small frequencies we may, as usually, expand our result by powers of $j\omega \tau$. Carrying terms up to order $(j\omega \tau)^2$, we obtain

\begin{align*}
J/J_0 & = 1 - j\omega \tau/2 - (\omega \tau)^2/6 = 1 - j\omega \tau/2 ; \\
|J/J_0|^2 & = 1 - 2(\omega \tau)^2/6 + (\omega \tau)^2/4 = 1 - (\omega \tau)^2/12 ; \\
|J/J_0| & = 1 - (\omega \tau)^2/24 .
\end{align*}

\((\ast 13-30)\)

\((\ast 13-30)\)

\((\ast 13-31)\)
5.2.2) The Richardson Equation.

We treat electrons as particles in a cube-shaped box of linear size $L$, and volume $V = L^3$. Quantum-mechanically, the electrons are plane waves with a wave vector $\mathbf{k}$. Periodic boundary conditions demand that for each of the components of $\mathbf{k}$

$$kL = 2\pi N,$$

where $N$ is an integer. The separation between adjacent $k$'s is therefore

$$\Delta k = \frac{2\pi}{L}.$$ (12)

The kinetic energy of the electron is

$$E = E_x + E_y + E_z,$$ (13)

where we have, for the contribution due to motion in the $x$–direction,

$$E_x = \frac{\hbar^2 k^2}{2m}.$$ (14)

with analogous relations for $E_y$ and $E_z$. At total energies sufficiently far above the Fermi level, the total number of electrons in a volume element $(dk)^3$ of $k$–space is

$$dN = 2 \cdot \exp\left[\frac{E - E_F}{kT}\right] \cdot \frac{dk_xdk_ydk_z}{(\Delta k)^3}.$$ (15)

Here, the leading factor 2 is due to spin. The current density contributed by the electrons in (15) is simply

$$dJ_z = -q v_x dN/L^3.$$ (16)

if $k_z > 0$ and $E_z > E_F$, and zero otherwise. Integration over all $\mathbf{k}$–values that satisfy these conditions yields, after some manipulation,
\[ J_z = \frac{2q}{(2\pi)^3} \cdot I_x \cdot I_y \cdot I_z \cdot \exp \left[ + \frac{E_F}{kT} \right], \]  
\[ \text{(17)} \]

where \( I_x, I_y, I_z \) are three integrals, given by

\[ I_x = \int_{k_x} \exp \left[ \frac{\hbar^2 k_x^2}{2mk_x T} \right] dk_x = \frac{\sqrt{2mk_x T}}{\hbar}, \]  
\[ \text{(18)} \]

\[ I_y = \int_{k_y} \exp \left[ \frac{\hbar^2 k_y^2}{2m' k_y T} \right] \frac{\hbar k_y}{m'} dk_y = \frac{kT}{\hbar} \cdot \exp \left[ \frac{E_a}{kT} \right]. \]  
\[ \text{(19)} \]

In the last integral, the lower limit \( k_y \) and the barrier height \( E_a \) are related via

\[ E_a = \frac{\hbar^2 k_y^2}{2m'}. \]  
\[ \text{(20)} \]

Putting it all together, we obtain the famous **Richardson Equation**

\[ J = -AT^2 \exp \left[ -\frac{q\Phi_B}{kT} \right], \]  
\[ \text{(21)} \]

where

\[ A = \frac{4\pi q m k^2}{(2\pi \hbar)^3} \equiv 120 \frac{A}{cm^2K^2} \frac{m}{m^*} \]  
\[ \text{(22)} \]

is the **effective Richardson constant** for a semiconductor with effective mass \( m^* \), and

\[ q\Phi_B = E_a - E_F \]  
\[ \text{(23)} \]

is the Schottky barrier height, measured from the Fermi level of the metal.
5.2.3) Arrhenius Plots and their Pitfalls.

The barrier height $q\Phi_b$ may be determined from experimental current density-vs. temperature data, by plotting the logarithm of the ratio $J/T^2$ as a function of $1/T$. According to (22), such a plot, called an Arrhenius plot of $J/T^2$, should yield a straight line with a slope $q\Phi_b/k$, and an intercept with the vertical axis at the logarithm of the constant $A$ in (22). If one does this, one often finds a quite different intercept. The principle reason for the deviation is that the barrier height tends to be slightly temperature-dependent. Suppose we have a linear dependence of $q\Phi_b$ over the temperature range of the data,

$$q\Phi_b = q\Phi_{b0} - \alpha \cdot kT. \quad (25)$$

Insertion of this form into (22) yields

$$J = -Ae^{\alpha^2} T^2 \exp\left[\frac{q\Phi_{b0}}{kT}\right]. \quad (26)$$

This is of the same form as (22), except that the constant $A$ has been replaced by $Ae^{\alpha^2}$, and the slope of the Arrhenius plot is not given by the actual barrier height anywhere within the temperature range of the data, but by the value linearly extrapolated to $T = 0$. Note that this value may differ significantly from the actual barrier height at $T = 0$, because the temperature dependence tends to become weaker as $T \to 0$.

Similar precautions apply to all forms of Arrhenius plots of quantities whose temperature dependence is dominated by a Boltzmann factor.

5.3) Image Force Barrier Lowering at Schottky Barriers.

The barrier height of Schottky barriers depends significantly on the electric field strength $E$ at the tip of the barrier, due to the effect of the electrostatic image force. When an electron is placed a distance $x$ ahead of a metal surface, an image charge is induced on the metal surface, which attracts the electron towards the metal, with the image force. It is an exercise in elementary electrostatics to show that the image force is given by
\[ F = \frac{q^2}{16\pi \varepsilon x^2}. \]  

(31)

where \( \varepsilon \) is the permittivity of the semiconductor. This force corresponds to an attractive potential energy, relative to the potential energy at infinity, of

\[ E(x) = \frac{q^2}{16\pi \varepsilon x}. \]  

(32)

To this potential energy we must add the true electrostatic potential energy \( -q|E|x \) (Fig. 1).

The sum of the two potential energies has a maximum at the position

\[ x = \sqrt{\frac{q}{16\pi \varepsilon E}}. \]  

(33)

which is typically of the order \( 10^{-7} \) cm. At the potential maximum, the potential energy is lowered relative to the potential without image force by the amount

\[ \Delta \Phi_B = \sqrt{\frac{qE}{4\pi \varepsilon}}. \]  

(34)
Fig. 1 Potential energy diagram ahead of a metal surface. The metal work function is \( q\Phi_m \). The potential barrier is lowered when an electrical field is applied to the surface, due to the combined effects of the field and the image force.

For an electric field of, say \( 100\,\text{kV/cm} \), and a dielectric constant of 10, one finds \( 37.95\,\text{mV} \), implying the scaling relation

\[
\Delta \Phi_g = 37.95 \, \text{mV} \cdot \sqrt{\frac{10q\Phi_m}{\varepsilon}} \cdot \frac{E}{100\,\text{kV/cm}}.
\]  

(35)

Evidently, with increasing reverse bias voltage the electric field at the interface increases, which pulls the barrier down (Figure 6.3-2), thereby increasing the reverse current.

Because the distance of the potential maximum from the metal is so small, an electron with a typical thermal velocity of about \( 10^7 \,\text{cm/s} \), the electron transverses the barrier in a very short time, of the order \( 10^{-14} \,\text{sec} \). The permittivity \( \varepsilon \) to be used in (34) and (35) is therefore not the same static permittivity, but the high-frequency permittivity for frequencies of the order \( 10^{14} \,\text{Hz} \). These are infrared frequencies, corresponding to wavelengths of about 3\( \mu \text{m} \). However, for most semiconductors, the infrared permittivities remain close to the static permittivities up to the energy gap equivalent frequency \( E_g / h \), which is usually larger than \( 10^{16} \,\text{sec} \). Hence, the dynamic corrections to the image force tend to be small. Fig. 3 gives an example, for Si.
Velocity Saturation Effects

We consider electron flow in only one direction, parallel or anti-parallel to the electric field $E$. We assume that, whenever the electron reaches a certain maximal speed $v_m$, it is immediately scattered, losing a certain amount of $\Delta E$ of kinetic energy in the process. As a result, its speed will drop to a value $v_s < v_m$. Its velocity after the scattering may be either $+v_s$ for a forward scattering event, or $-v_s$ for backward scattering. Inelastic scattering process tends to be isotropic, so we make the assumption that forward and backward scattering are equally probable. We are only interested in the time-averaged electron velocity; for equal scattering probabilities it may be calculated as if forward and backward scattering events alternated, rather than being randomly distributed. For alternating forward and backward scattering, the electron velocity as a function of time behaves as shown in Fig. 3.

Following each scattering event, the electron is accelerated from the velocity $\pm v_s$ to the velocity $v_m$, at which the point the next scattering event takes place. Following a forward scattering event, the electron reaches the velocity after a time interval...
\[ t_+ = (v_m - v_e) \cdot m' / qE, \]  
\[ (13-7a) \]

If no other kinds of scattering events intervene. Following a backward scattering event, the electron is first decelerated, and then accelerated again, reaching the velocity \( v_m \) after a time

\[ t_- = (v_m + v_e) \cdot m' / qE, \]  
\[ (13-7b) \]

again assuming that no other kinds of scattering events intervene.

The qualifiers with regard to other kinds of scattering events are important. But we note from \( (13-7a,b) \) that the times to the next inelastic scattering event decrease with increasing field strength, thereby decreasing the probability of another kind of scattering event intervening. At sufficiently large fields, such alternate events become negligible. The average velocity in that limit, called the saturated drift velocity, is then easily found to be

\[ v_s = (v_m^2 - v_e^2) / 2v_m, \]  
\[ (13-8) \]

independent of the electrical field strength! Although the rate of acceleration increases with increasing field strength, the rate of velocity-reducing collisions increases in the same way, and the field strength cancels out of the average velocity.

We is often useful to write \( (13-8) \) in an alternate way,

\[ v_s = v_m \cdot 2E_m, \]  
\[ (13-9) \]

where \( E_m \) is the kinetic energy associated with the velocity \( v_m \).

Although highly oversimplified, our treatment captures the essence of the physics of drift velocity saturation. It is readily possible to refine the model, but the mathematical effort soon becomes prohibitive, without a major change in the final result.

In our treatment, we neglected the near-elastic scattering that dominates the electron dynamics at low electric fields; hence this treatment describes only the high-field limit. At low fields, we have, of course the simple linear relation

\[ v_0 = \mu_e E. \]  
\[ (13-10) \]

Somewhere in the vicinity of the critical field

\[ E = E_c = v / \mu_e \]  
\[ (13-11) \]
the two relationships cross over. A mathematical theory that would cover the crossover range as well as would be a difficult undertaking. For purposes of device physics calculations it is usually sufficient to somehow interpolate between the two relations. Two simple interpolations are widely used:

\[ t^* = (v_m - v_r) \frac{m^*}{qE} \]

from \( a = \frac{F}{m} = \frac{qE}{a} \)

and \( v = u + at \)

\[ t^* = (v_m - v_r) \frac{m^*}{qE} \]

\[ t^* - t_o = v_m \cdot a \]

\[ v_{sr} = \frac{1}{T} \left[ -v_r \cdot \frac{t_o}{2} + v_m \frac{(t^* - t_o)}{2} + (v_m - v_r) \frac{t^*}{2} + v_r \cdot t^* \right] \]

\[ v_{sr} = \frac{1}{T} \left[ -v_r \cdot \frac{a}{2} + v_r \cdot a + \frac{(v_m - v_r)^2}{2} \cdot a + v_r \cdot (v_m - v_r) \right] \]
13.4) Space Charge Effects in the Base-Collector Depletion Layer
(Kirk Effect)

One of the consequences of drift velocity saturation is that it is not possible to speed up the current transport through the base-to-collector depletion layer of a bipolar transistor by simply increasing the collector reverse bias. At high current densities that leads to appreciable space charge effects inside the depletion layer, and there will be a certain maximal current density beyond which the collector ceases to collect carriers effectively.

We assume that the base/collector junction is a $p^+n^-n^+$ junction, with an $n^-$ region of width $w^-$. In the absence of any current, the electric field distribution within the depletion layer will be as in line “a” in Fig.4, with the maximal field occurring at the base end of the depletion layer: With increasing current this distribution first flattens out and then reverses; for constant reverse bias, the area under the line remains constant, too. At a current density

$$J_c = q n^- \cdot v_i$$

the space charge of the traveling current exactly cancels that of the stationary donors, and the field becomes uniform (line “b”). For higher current densities the field distribution reverses. Eventually, the electric field at the base end of the depletion region will go to zero (line “c”); this happens when the current density reaches the value
When the current density exceeds this value, the overall result is the drastic extension of the effective base width into the region between the metallurgical base and the collector (Kirk Effect).

\[ J_c = J_{\text{max}} = \left( \frac{2e V_{\text{bi}}}{w^*} + qn \right) v_x. \]
TWO INTEGRAL RELATIONS PERTAINING TO THE ELECTRON
TRANSPORT THROUGH A BIPOLAR TRANSISTOR WITH A
NONUNIFORM ENERGY GAP IN THE BASE REGION

H. KROEMER
Department Of Electrical and Computer Engineering, University of California, Santa Barbara,
CA 93106, U.S.A

(Received 23 November 1984; revised form 23 February 1985)

Abstract: The two integral relations by Moll and Ross for the current flow through the base
region of a bipolar transistor, and for the base transit time, are generalized to the case of a
heterostructure bipolar transistor with a nonuniform energy gap in the base region.

The recent strong interest in
heterostructure bipolar transistors
(HBTs) with a nonuniform energy gap in
the base region [1] calls for an extension
of bipolar transistor theory to transistors
that exhibit such a variation. The first
priority is a generalization of the familiar
integral relation for the current flow
through the base region,

\[ J_x = \frac{q D_n n_i^2 \times \exp(qV_{BE}/kT)}{\int_B N_x dx} \]  

(1)

This relation was first derived in slightly
different form by Moll and Ross [2]; it
forms the backbone of the Gummel-
Poon model [3] wisely used in transistor
modeling, for example in the SPICE
computer program [4].

According to (1), the current density
for a given base-emitter voltage
depends only on the total number of
acceptors per unit area in the base, the
so-called Gummel Number, but not on
the spatial distribution of these
acceptors. As shall we shall see, this
changes drastically in a graded-gap
structure.

A second quantity of interest is the
electron transit time through the base
region. It does depend on the acceptor
distribution; Moll and Ross [2] first
derived the relation

\[ \tau = \frac{1}{D_n} \int_B \frac{1}{N_x} \left[ \int_y N_y dy \right] dx. \]  

(2)

A very "physical" rederivation and
discussion of this relation was
subsequently given by Varnerin [6].

The relations (1) and (2) are for the
simplest case, when a number of
approximations may be made, such as
low-level injection, position-independent
diffusivity, negligible hole current, and
the absence of hot-electron effects
causing nonlinearity in the basic drift-
and-diffusion transport relations.

Refinements of the current flow relation
(1) that remove the various
approximations, but which retain the
uniform energy gap, have been
extensively discussed in the literature [3-
5]. The transit time relation (2),
however, has been largely ignored by
subsequent workers [7].

Our point of departure is the set of
linear transport relations for the electron
and hole contribution to the electrical
current density across the base region of
an npn-transistor, expressed in their most
general form, involving the two carrier
concentrations and the gradients of the
two quasi-fermi levels $E_n$ and $E_p$ for electrons and holes:

$$J_n = +\mu_n p \frac{d}{dx} \left( \frac{E_n}{q} \right) = +\mu_n n \frac{d}{dx} E_n,$$

(3a)

$$J_p = -\mu_p p \frac{d}{dx} \left( \frac{E_p}{q} \right) = -\mu_p p \frac{d}{dx} E_p.$$

(3b)

This formulation implies the neglect of hot-electron effects. In a good transistor, the hole current may also be neglected, which implies $0$. Integration of (7) from an arbitrary point $x$ within the base region to the collector edge $x = w$ yields

$$np \left[ \frac{n}{n_i^2} \right]_w^0 = -np \left[ \frac{n}{n_i^2} \right]_w^0 = J_s \int_w^0 \frac{p}{D_n n_i^2} dy.$$

(8)

In the first equality we have assumed that the collector is reverse-biased, and that $np \left[ \frac{n}{n_i^2} \right]_w^0$ may therefore be neglected. At the emitter end of the base, $x = 0$, we apply Shockley boundary conditions:

$$np \left[ \frac{n}{n_i^2} \right]_w^0 = \exp \left( qV_{be}/kT \right).$$

(9)

If (9) is inserted into (8), and the result solved for $J_n$, we obtain the desired generalization of the Moll-Ross current relation:

$$J_n = -q \times \exp \left( qV_{be}/kT \right) \int_0^w \left[ \frac{p}{D_n n_i^2} \right] dy.$$

(10)

We return to (8) for arbitrary $x$, and write it in the form

$$n(x) = -(J_s/q) \times (n_i^2/p) \times \int_w^x \left[ \frac{p}{D_n n_i^2} \right] dy.$$

(11)

Integration over the base and division by $-J_s/q$ yields the average transit time through the base:

$$\frac{q}{J_s} \int_0^w n(x) dx \left[ \frac{p}{D_n n_i^2} \right] d y = \frac{\int_0^w \left[ (n_i^2/p) \int_w^y \left[ \frac{p}{D_n n_i^2} \right] d y \right] dz}{J_n}.$$

(12)
This is the desired generalization of the transit time relation (2). In the uniform-base case, the new relations (10) and (12) both reduce to their already known limits (1) and (2), as they must.

The relation s (10) and (12) are the two central results of this paper.

As stated earlier, HBTs are likely to have base doping levels much higher than homostructure transistors. In fact, in those HBTs that are of practical interest, the doping levels in the base will be higher than in the emitter. As a result, the various kinds of high-level injection effects in which the electron concentration in the base rises significantly above the doping concentration, and which play such an important role in homostructure transistor theory, cannot occur in HBTs of practical interest, and may therefore be ignored. Hence, it will, as a rule, be an excellent approximation to replace the hole concentration \( p \) in (10) and (12) by the acceptor doping concentration \( N_A \) even at the highest current densities the emitter is capable of delivering.

It is evident from (10) and (12) that in a variable-gap transistor the acceptor doping concentration is not weighted uniformly throughout the base, as in the uniform-gap case, but with the weighting factor \( I(n^2) \), which in turn is proportional to \( \exp(E_x/kT) \).

To illustrate the application of the new relations, especially of the transit time relation (12), it is useful to discuss, as an instructive example, the case of a linear variation of the energy gap with position,

\[
E_x = E_{x0} - qFx = E_{x0} - \Delta E_x(x/w), \tag{13}
\]

with all other parameters position-independent, and also assuming \( p = N_A \).

Evidently, the quantity \( F \) in (13) is the built-in drift field for the electrons, and \( \Delta E_x \) is the total variation of the energy gap within the base region. Equation (13) implies

\[
n^2_{i,w} = n^2_{i,0} \times \exp\left(qFx/kT\right), \tag{14}
\]

where the notation \( n_i \) refers to the values of \( n_i \) taken at the position \( x \). If (14) is inserted into (10), the result may be written

\[
J_n = \frac{q^2 F D_n}{N_A kT} \left( \frac{n_{i,0} n_{i,0}^2}{n^2_{i,w} - n^2_{i,0}} \right) \times \exp\left(qV_{BE}/kT\right). \tag{15}
\]

Note that, if emitter and collector are interchanged, the magnitude of the current remains unchanged; only the sign changes. This is not the case for the transit time: insertion of (14) into (12) yields, after some manipulation,

\[
\tau = \left( w/\mu_e F \right) \times \left[ 1 - \left( kT/qFw \right) \times \left[ 1 - \left( n_{i,0}/n_{i,w} \right)^2 \right] \right]. \tag{16}
\]

For normal transistor operation we have \( n_{i,0}/n_{i,w} \ll 1 \), which case (16) may be simplified to

\[
\tau = \left( w/\mu_e F \right) \times \left[ 1 - \left( kT/qFw \right) \right]
= w/\mu_e F = \tau_0 \times \left( 2kT/\Delta E_x \right) \ll \tau_0, \tag{17}
\]

where, on the second line, we have introduced, for comparison, the base transit time for a uniform-base transistor,
\[ \tau_o = \frac{w^2}{2D_u}, \quad (18) \]

and where we have assumed that \[ \Delta E_g \gg kT. \]

For inverted operation the inequality \[ \frac{n_i}{n_w} \gg 1 \] holds, and we find

\[ \tau = \left( \frac{w}{\mu} \right) \times \left\{ 1 + \left( \frac{kT}{qF_w} \right) \times \left( \frac{n_i}{n_w} \right)^2 \right\} \]

\[ = \frac{2\tau_o}{\times} \times \left( \frac{kT}{\Delta E_g} \right)^2 \times \left( \frac{n_i}{n_w} \right)^2 \gg \tau_o. \quad (19) \]

The asymmetry in the forward and reverse transit times implied in (17) and (19) is self-evident. The results (16)-(19) for the specific case of a uniformly doped base with a linearly varying gap are identical to the expressions for a uniform-gap transistor in which a uniform electric field has been incorporated into the base region by means of an exponential doping profile, a case widely discussed during the 1950s [8,2,6]. The simple limit (17) was recently restated by Hayes et al. [1].

What is new is the generalization of (2) to essentially arbitrary variations of the energy gap, in addition to that of other base parameters.

Acknowledgements—I wish to express my thanks to Dr. Peter Asbeck for helpful discussions, and to Prof. Fred Lindholm and Dr. James M. Early for helping me find some of the old literature on that subject.

REFERENCES


4. See, for example, I. Getreu, Modeling the Bipolar Transistor, Tektronix Inc. (1976).


EARLY VOLTAGE

(Physical picture)

\[ \frac{\partial Q_b}{\partial V_{cb}} = C_{bc} = \frac{\varepsilon}{W_c} \]

\[ Q_b = Q_{b0} - \frac{\varepsilon}{W_c} \Delta V_{cb} \quad \text{for constant base doping} \]

\[ x_b = x_{b0} - \frac{\varepsilon}{W_c} \frac{\Delta V_{cb}}{qN_A} \]

\[ J_c - J_c(\Delta V_{cb}) = qD_n \frac{n_{m}^{\frac{1}{3}}}{x} - \frac{k}{x} \]

\[ J_c = \frac{k}{x_{b0} - \frac{\varepsilon}{W_c} \frac{1}{qN_A} \Delta V_{cb}} \]

when the denominator \( \rightarrow 0 \) \( J_c \rightarrow \infty \) PUNCHTHROUGH

\[ J_c = \frac{k}{x_{b0}} \frac{1}{1 - \frac{\varepsilon}{W_c} \frac{1}{qN_A x_{b0}} \Delta V_{cb}} = \frac{k}{x_{b0}} \left( 1 + \frac{C_{bc}}{Q_{b0}} \Delta V_{cb} \right) \]

\[ \Delta J_c = \frac{k}{x_{b0}} \frac{\Delta V_{cb}}{V_A} \text{ or } \frac{\Delta V_{cb}}{\Delta J_c} = \frac{V_A}{I_c} \]
specifically considered thus far. For example, a transistor as described in Chapter 6 would have a current-source output in the active mode; that is, output current would not depend in any way on the voltage applied between the base and collector terminals. In real transistors, however, both output currents and output voltages influence the device performance. Similarly, the theory that has thus far been presented would predict a current gain in the active mode that is insensitive to bias level. But this is only a rough approximation. A first objective of this chapter, therefore, is to investigate mechanisms that must be considered to obtain a more realistic picture of transistor operation.

Another major inadequacy of the theory developed thus far is that it does not deal with time-varying effects. Their consideration is therefore the second major topic to be developed in this chapter. A great simplification in considering time-varying effects will be obtained by modeling of transistors as charge-controlled devices. This viewpoint leads very naturally into the important topic of large-signal transient models for bipolar transistors. The amalgamation of the Ebers-Moll model with the charge-control model provides a powerful analysis technique. It will be straightforward to derive a small-signal ac equivalent circuit from the more general large-signal representation. This small-signal equivalent circuit, known as the hybrid-pi circuit, is relatively easy to characterize because it is composed of elements that relate directly to physical mechanisms in the transistor.

The production of pnp transistors within the confines of standard IC processing is described in a final section. Two basic types of pnp transistors, substrate pnp and lateral pnp transistors are typically fabricated. The substrate pnp transistor has only limited application because its collector is not isolated. The lateral pnp transistor has severely limited performance compared to that of npn transistors. Simple models for loss mechanisms in this device explain the limited performance. An important application of lateral pnp transistors—a dense form of logic circuitry known as integrated-injection logic (I^2L)—is described at the conclusion of the chapter.

7.1 Effects of Collector Bias Variation (Early Effect)

When we considered bipolar transistors under active bias in Chapter 6, the function of the voltage applied across the collector-base junction was merely to insure the efficient collection of base minority carriers and their delivery to the collector region. The magnitude of the bias only limited the range of the permissible collector voltage swing, provided that it was below the breakdown value. We have overlooked an important aspect of pn-junction electronics in taking this simplified view. As we noted in Chapter 4, the width of a reverse-biased pn junction is voltage dependent; in fact, this bias dependence made possible the operation of junction field-effect transistors. In the case of bipolar transistors, a changing collector-base bias causes variation in the space-charge layer width at the collector junction and, consequently, in the width of the quasi-neutral base region. These variations result
in several effects that complicate the performance of the device as a linear amplifier. Base-width modulation as a consequence of variations in collector-base bias was first analyzed by James Early,3 and the phenomenon is generally called the Early effect.

The dependence of collector current on collector-base bias can be formulated directly by using the integral equations for active bias (npn transistor) developed in Section 6.2. In particular, from Equations 6.2.1 and 6.1.10, we can write

\[ I_c = \frac{qD_s \eta_i^2 A_e \exp(qV_{BE}/kT)}{\int_0^{x_B} p \, dx} \]  (7.1.1)

where the integration is performed over \( x_B \), the width of the quasi-neutral base region and the other terms have been defined in Section 6.1.

Variation in the base width with voltage \( V_{CB} \) causes a variation in the collector current that can be written

\[ \frac{\partial I_c}{\partial V_{CB}} = -\frac{qD_s \eta_i^2 A_e \exp(qV_{BE}/kT)p(x_B)}{\left[ \int_0^{x_B} p \, dx \right]^2} \frac{\partial x_B}{\partial V_{CB}} \]  (7.1.2)

Several of the terms in Equation 7.1.2 can be combined to represent collector current itself, so that Equation 7.1.2, which represents the small-signal conductance at the collector-base junction, can be written

\[ \frac{\partial I_c}{\partial V_{CB}} = -I_c \left[ \frac{1}{\int_0^{x_B} p \, dx} \right] \frac{\partial x_B}{\partial V_{CB}} \]  

\[ = -\frac{I_c}{V_A} = \frac{I_c}{|V_A|} \]  (7.1.3)

Since the collector is reverse biased, the derivative in Equation 7.1.3, \( \partial x_B/\partial V_{CB} \), is negative, and the Early effect results in an increase in \( I_c \) when \( V_{CB} \) is increased. This increase is evident if we examine common-emitter output characteristic curves such as the family shown in Figure 7.1a. These curves are actually plots of \( I_c \) versus \( V_{CE} \), but \( V_{CE} \) has almost the same value as \( V_{CB} \) for bias in the active region (\( V_{CE} \approx V_{CB} + 0.7 \) V).

Equation 7.1.3 reveals that the Early effect varies linearly with collector current. The reciprocal of the factor multiplying the current has the dimension of a voltage and has been defined as the Early voltage. The Early voltage is usually given the symbol \( V_A \). From Equation 7.1.3, an equation for \( V_A \) for an npn transistor is

\[ V_A = \frac{\int_0^{x_B} p \, dx}{p(x_B)\partial x_B/\partial V_{CB}} \]  (7.1.4)
Figure 7.1  Measured output characteristics for an amplifying transistor:
collector current versus \( V_{CE} \). (a) vertical 0.1 mA/div., horizontal 1 V/div.
(b) vertical 0.1 mA/div., horizontal 10 V/div.; tangents to the measured curves
(at the edge of saturation) are extended to the voltage axis (dashed lines) to
determine the Early Voltage \( V_A \). Extension of lines drawn approximately
tangential to the characteristic in the active region intersects the voltage axis
at \( V_A' \) (solid lines).

Again, the derivative in Equation 7.1.4 is negative and, thus, the Early voltage is
negative for an npn transistor. The analogous effect of emitter-base space-charge
layer widening when the transistor is biased in the reverse-active mode can also be
interpreted by using a different Early voltage, usually denoted as \( V_B \).

Except for high-level effects, the three terms that define \( V_A \) depend only on the
transistor manufacturing process and on collector-base voltage. In practice, the
collector-base voltage dependence of \( V_A \) itself is usually treated as negligible and
the Early voltage is approximated by its value at a single bias, often at \( V_{CB} = 0 \).
By using this condition to specify \( V_A \), we can expect that a series of tangents to
curves of \( I_C \) versus \( V_{CB} \) (\( V_{CE} \) in practice), drawn at the edge of the forward-active
region where \( V_{CB} \) is approximately zero, should intersect the \( V_{CE} \) axis at \( V_A \). In
Figure 7.1b dashed lines drawn from the point at which the transistor moves out
of saturation do intersect at a common point, indicating the Early voltage. For
circuit-design and analysis, however, interest is not in an Early voltage characteri-
zizing the edge of saturation, but rather in a parameter to use in the forward-active
region. If tangents are drawn to the curves of \( I_C \) versus \( V_{CE} \) in the active
region, they do not, in general, intersect each other on the voltage axis. It is usual,
however, to approximate an intersection point appropriate to the range of bias of the
transistor as at \( V_A' \), formed by the solid lines in Figure 7.1. This construction
forms similar triangles \( AOB \) and \( BCD \) (if we neglect the small saturation-voltage
offset of the BJT curves). We can infer Equation 7.1.3 directly from these similar
triangles if the differential elements \( dI_C \) and \( dV_{CB} \) in Equation 7.1.3 are treated as
incrementals \( \Delta I_C \) (indicated by line \( CD \)) and \( \Delta V_{CB} \) (approximately indicated by line
\( BC \)) since line \( AO \) indicates \( V_A' \) and line \( OB \) indicates \( I_C \).
7.1 Effects of Collector Bias Variation (Early Effect)

A useful and informative alternative expression for $V_A$ can be obtained by rearranging some terms in Equation 7.1.4. First, we use Equation 6.1.8 to express the numerator in terms of the base majority-charge density $Q_b$ in that portion of the base where transistor action is taking place:

$$\int_0^{x_b} p \, dx = \frac{Q_b}{q}$$  \hspace{1cm} (7.1.5)

We then recognize that the denominator of Equation 7.1.4 represents the derivative of base charge $Q_b$ with respect to $V_{CB}$:

$$q p(x_b) \frac{dx_b}{dV_{CB}} = \frac{dQ_b}{dV_{CB}}$$  \hspace{1cm} (7.1.6)

The derivative on the right-hand side of Equation 7.1.6 can be related to $C_{je}$, the small-signal capacitance per unit area at the collector-base junction:

$$\left| \frac{dQ_b}{dV_{CB}} \right| = C_{je}$$  \hspace{1cm} (7.1.7)

Therefore, the Early voltage is

$$|V_A| = \frac{Q_b}{C_{je}}$$  \hspace{1cm} (7.1.8)

To reduce the influence of collector-base voltage on collector voltage, $V_A$ should be increased in magnitude. From Equation 7.1.8 we see that this can be accomplished in practice by increasing the ratio of base majority-charge per unit area to the capacitance per unit area at the collector-base junction. Physically, this reduces the movement of the base-collector boundary into the base region.

It is useful, for computer modeling purposes, to describe the widening of the base-collector space-charge layer through the Early voltage $V_A$, as we shall see in Section 7.6. It may, however, help conceptually to consider an alternative view of the Early effect that is physically more revealing for the prototype transistor introduced at the beginning of Chapter 6. For homogeneous doping and active-mode bias, the minority-carrier distribution in the base has a triangular shape (Figure 7.2). Increasing the collector-base voltage from $V_{CB}$ to $(V_{CB} + \Delta V_{CB})$ moves the base-collector junction edge a distance $\Delta x_b$ from $x_{80}$. A second triangular distribution specifies the new minority-carrier profile, and the shaded area between the two distributions represents the decrease in stored base charge.

The collector current is thus increased in the ratio $x_{80}/x_{81}$ by the change in $V_{CB}$. In addition, it becomes clear by looking directly at stored base charge that the Early effect reduces charge storage. This affects both the transient behavior of transistors and the dc base current since base recombination depends directly on stored base charge. We shall consider these effects in more detail in Section 7.5 when we discuss BJT models for circuit applications.
maintain acceptable injection efficiency [which constrains \( N_d(0) \) to be appreciably lower than the emitter doping] and the requirement of an extrinsic p-type doped material at \( x = x_b \) [which forces \( N_d(x_b) \) to be greater than \( N_{d_{epn}} \)].

### 7.4 Charge-Control Model

A framework for bipolar-transistor equations that is especially useful for time-dependent analysis is called the charge-control model. In this model the controlled variable is not current or voltage; instead, equations are framed in terms of controlled charges within regions of the device.

A typical charge-control relationship for a transistor under active bias was derived in the previous section in Equation 7.3.2. This equation, \( I_C = \frac{Q_{ab}}{\tau_B} \), relates the minority charge stored in the quasi-neutral base \( Q_{ab} \) to the current \( I_C \) carried by transistor action between the emitter and the collector. The charge and current are linearly related with \( \tau_B \), the transit time in the quasi-neutral base, as the proportionality factor. Because Equation 7.3.2 represents only minority-carrier transport across the base, however, it is just one portion of the charge-control model for a transistor.

Control in an amplifying npn transistor is exercised by the bias on the base-emitter junction. This bias affects not only \( Q_{ab} \), but other charge components as well. The major additional components to be considered are the charges represented by holes injected into the emitter, which we designate as \( Q_{pe} \), and the charges stored on the base-emitter and base-collector depletion capacitances, which are given the symbols \( Q_{ve} \) and \( Q_{vc} \), respectively. Figure 7.14 shows these components for a prototype transistor. We first discuss the two injection components, \( Q_{ab} \) and \( Q_{pe} \) that are responsible for steady-state base current. The other charge components shown in Fig. 7.5 and will be considered in the emitter voltage is increased when the BJT is the controlling (base major negative for a npn transistor) terms of \( Q_F \) if a character analogy to Equation 7.3.2

Note that all that is known \( I_C \) must be linearly related to \( Q_F \).

Because \( Q_F \) represents the quasi-neutral emitter as that of a diode, and we may

where \( Q_{pe} \) is a function of at least somewhat greater \( \tau_B \) (Equation 7.3.2). Further control model is developed

The steady-state current which \( Q_{ab} \) recombines in the base and injected into the emitter to a diode factor \( \exp(qV_{be}/kT) \) to \( Q_F \). It is therefore possible to express the gain

A considerable amount of recombination processes is required to express \( \tau_B \)

analysis is not of special value from measurements, and representing the device for design models.

Equations 7.4.1 and 7.4.4 gain is simply the ratio of
components shown in Figure 7.14 influence the time-varying behavior of the BJT and will be considered later. Since both \( Q_{nb} \) and \( Q_{pe} \) increase when the base-emitter voltage is increased, their sum \( (Q_{nb} + Q_{pe}) \) is called \( Q_F \) (because \( Q_F \) is increased when the BJT is under forward-active bias). The sign of \( Q_F \) is the sign of the controlling (base majority-carrier) charge, positive for an n-p-n transistor and negative for a p-n-p transistor. The steady-state collector current can be written in terms of \( Q_F \) if a characteristic time \( \tau_F \) is introduced. The equation for current (in analogy to Equation 7.3.2) is

\[
I_C = \frac{Q_F}{\tau_F}
\]  

(7.4.1)

Note that all that is formally required to make Equation 7.4.1 accurate is that \( I_C \) must be linearly related to \( Q_F \) (with \( \tau_F \) taken to be constant).

Because \( Q_F \) represents the sum of the magnitudes of the excess minority charge in the quasi-neutral emitter and base regions, its voltage dependence is generally that of a diode, and we may write

\[
Q_F = Q_{FO} \left[ \exp \left( \frac{qV_{BE}}{K T} \right) - 1 \right]
\]  

(7.4.2)

where \( Q_{FO} \) is a function of the dopant profiles and device geometry. Since \( Q_F \) is at least somewhat greater than \( Q_{nb} \), \( \tau_F \) must be greater than the base transit time \( \tau_B \) (Equation 7.3.2). Further comments about \( \tau_F \) are deferred until the charge-control model is developed more fully.

The steady-state current flowing in the base lead is proportional to the rate at which \( Q_{nb} \) recombines in the quasi-neutral base plus the rate at which holes are injected into the emitter to replenish \( Q_{pe} \). These two rates are proportional to the diode factor \( \exp(qV_{BE}/K T) - 1 \), and therefore (by Equation 7.4.2) proportional to \( Q_F \). It is therefore possible to write a charge-control expression for the input (base) current of the transistor

\[
I_B = \frac{Q_F}{\tau_{BF}}
\]  

(7.4.3)

A considerable amount of physical analysis, involving emitter efficiency and the recombination processes for excess carriers in both the emitter and base, would be required to express \( \tau_{BF} \) or \( \tau_F \) in terms of more fundamental parameters. This analysis is not of special value because both \( \tau_{BF} \) and \( \tau_F \) can be obtained in practice from measurements, and also because the charge-control model aims at representing the device for design, not at exploring the physical electronics of transistor models.

Equations 7.4.1 and 7.4.3 can be used to show that the steady-state current gain is simply the ratio of the two characteristic times:

\[
\frac{I_C}{I_B} = \beta_F = \frac{\tau_{BF}}{\tau_F}
\]  

(7.4.4)
For example, Equations 7.4.1, 7.4.3 and 7.4.4 can be applied to the prototype transistor of Figure 6.1. If the emitter efficiency is very high in the prototype device, then \( Q_E \approx Q_{EB} \) with \( Q_{EB} = \frac{1}{4} \mu n(0) x_B A E \). Since only base recombination is significant for this case, \( \tau_{BF} = \tau_e \) and \( \tau_f = x_B^2/2 \bar{D}_e \) as derived in Equation 7.3.3. Therefore, from Equation 7.4.4 \( \beta_f = 2 L_2 x_B^2 / x_B^2 \) where \( L_2 = \sqrt{\bar{D}_e} \tau_e \). This result for dc current gain can be compared with earlier analysis of the same problem in Section 6.2. There, \( \alpha_f \) was derived as the base transport factor to be \([1 - (x_B^2 / 2 L_2^2)]\) in Equation 6.2.8. If we use this result for \( \alpha_f \) in the equation \( \beta_f = \alpha_f / (1 - \alpha_f) \), we find an identical value for \( \beta_f \) to that obtained from the charge-control analysis.

A full charge-control model for the bipolar transistor is derived by adding terms to represent the currents that flow because of time variations in stored charge. Clearly, if \( Q_f \) increases with time, there will be a component of base current equal to \( dQ_f / dt \). Likewise, changes in the charges stored at the base-emitter and base-collector junctions (\( Q_{VE} \) and \( Q_{VC} \)) result in added base current. An overall expression for the base current is, therefore,

\[
i_b = \frac{Q_f}{\tau_{BF}} + \frac{dQ_f}{dt} + \frac{dQ_{VE}}{dt} + \frac{dQ_{VC}}{dt}
\]  

(7.4.5)

The first three components of current in Equation 7.4.5 flow from the base to the emitter; the last flows from the base to the collector. Combining Equation 7.4.1 and 7.4.5 and using Kirchhoff's current law, one obtains a set of charge-control equations for the transistor under active bias:

\[
i_C = \frac{Q_f}{\tau_f} - \frac{dQ_{VC}}{dt}
\]

\[
i_b = \frac{Q_f}{\tau_{BF}} + \frac{dQ_f}{dt} + \frac{dQ_{VE}}{dt} + \frac{dQ_{VC}}{dt}
\]

\[
i_E = -Q_f \left( \frac{1}{\tau_f} + \frac{1}{\tau_{BF}} \right) - \frac{dQ_f}{dt} - \frac{dQ_{VE}}{dt}
\]

(7.4.6)

We have thus derived a set of linear equations relating currents and charges in a bipolar transistor. These equations are in contrast to the nonlinear expressions that relate currents and voltages in the transistor.

The circuit diagram of Figure 7.15 represents the various terms in Equation 7.4.6. The diode from the base to the emitter passes the steady-state current and has a saturation current \( i_{ES} = Q_{F0} [(1/\tau_f) + (1/\tau_{BF})] \) as indicated by Equation 7.4.2. The elements storing the charges \( Q_f, Q_{VE}, \) and \( Q_{VC} \) are shown as capacitors with a line across them to indicate that they store charge (like capacitors), but are voltage dependent. Now that a basic set of charge-control equations has been written for a BJT, we are better able to gain a perspective on the limitations of this viewpoint.

The basic premise underlying charge-control analysis is the existence of a constant proportionality between quantity of charge and current. Another way of stating this is that the characteristic times in the charge-control equations must not themselves be functions of dynamic conditions, etc. these functions of the carriers must be functions of the base-emitter junction and not have the same ratio of these parameters.

If a dynamic problem for charge are constrained. Hence, these solutions may have greater part of the physical basis, but a fair representation of the significant error in the model. in the base, that is, the other parameters is appreciably greater than the perfect clear.

Applications of the Charge-Control Analysis

Before discussing the basic concepts, we illustrate the use of Equations 7.4.1 and 7.4.6 for this purpose is shown in Figure 7.16. The npn transistor under active bias is sketched in Figure 7.16. The transient calculations, severa can be considered negligible under forward bias. This is because we have chosen it is also possible to treat...
7.4 Charge-Control Model

The prototype transistor, the prototype device, the combination is significant for dynamic conditions. Therefore, the characteristic times may be derived for dc conditions and then be used to express dynamic conditions. Strictly this premise is not correct; the characteristic times are functions of the charge. For example, to affect current at the collector, minority carriers must disperse through the base region after being injected at the edge of the base-emitter junction. Thus, during transient conditions, collector current does not have the same ratio to base charge as it does in the steady state.

If a dynamic problem is analyzed with the charge-control model, the solutions for charge are constrained to be a time sequence of differing “steady-state” solutions. Hence, these solutions are sometimes called quasi-static approximations. For the greater part of the transient, however, the charge-control solution is usually a fair representation of the more exact result. The time scale over which there is significant error in the charge-control solution is of the order of the transit time in the base, that is, the order of $\tau_F$. For most applications the time scale of interest is appreciably greater than $\tau_F$, and solution by means of the charge-control model is perfectly adequate. A specific example in the next section will make this more clear.

Applications of the Charge-Control Model

Before discussing the bipolar charge-control model further, it is worthwhile to illustrate the use of Equations 7.4.6 in an application. A simple circuit appropriate for this purpose is shown in Figure 7.16a, and the circuit plus charge-control model is sketched in Figure 7.16b. What is sought is the collector-current response of an npn transistor under active bias when driven by a current source at the base. For hand calculations, several simplifications are appropriate. First, the current $dQ_{VC}/dt$ can be considered negligible since the base-emitter voltage varies only slightly under forward bias. This simplification is usually made for active-bias operation. Because we have chosen a simple circuit in which the collector voltage is constant, it is also possible to treat the current $dQ_{VC}/dt$ as negligible.
Figure 7.16 (a) Simple circuit for illustration of charge-control model. (b) Equivalent-circuit model. The cancelled elements carry negligible currents.

With these approximations the equation for the base current has only one unknown, the controlled charge $Q_F$, and takes the form

$$i_b = \frac{Q_F}{\tau_{BF}} + \frac{dQ_F}{dt}$$  \hspace{1cm} (7.4.7)

Since $i_b(t)$ is specified by

$$i_b = i_{b1}(t < 0) = i_{b2}(t > 0)$$

the solution for $Q_F$ will contain terms associated with both the homogeneous and particular forms of Equation 7.4.7. When the boundary values $Q_F(t = 0) = i_{b1}\tau_{BF}$ and $Q_F(t \to \infty) = i_{b2}\tau_{BF}$ are used, the solution becomes

$$Q_F = \tau_{BF}[i_{b2} + (i_{b1} - i_{b2}) \exp(-t/\tau_{BF})]$$  \hspace{1cm} (7.4.8)

Under the approximations that we have made, the collector current is given by $Q_F/\tau_F$ so that its time dependence becomes (using Equation 7.4.4)

$$i_c = \frac{\beta_F}{\tau_F}[i_{b2} + (i_{b1} - i_{b2}) \exp(-t/\tau_{BF})]$$  \hspace{1cm} (7.4.9)

The collector current thus changes from its initial to its final value following an exponential function with a characteristic time constant equal to $\tau_{BF}$ (Figure 7.17).

As mentioned in the previous section, the transient solution given in Equation 7.4.9 is in error for small values of $t$. At zero time, for example, Equation 7.4.9 predicts an abrupt change in the slope of the collector current equal to $(i_{b2} - i_{b1})/\tau_F$ whereas co-injected at the emitter side that takes account of the not change at all for time emitter currents are equ. reaches the base transit time transient solution predicts that result for times large.

The foregoing examples have been impo. approximations that have been avoided mathematical cor.

One such complication was that we were not connected to an through a load resistor $R_L$ required to charge $Q_{VC}$ would readily if we define an effect over the collector voltage $i$ to Equation 7.4.9 provided by [Problem 7.19]:

To derive Equation 7.4.10, it capacitance affects the base c
The homogeneous and inhomogeneous solutions are given by
\[ Q_f(t = 0) = i_{n1} \tau_B F \]
and
\[ Q_f(t = 0) = i_{r1} \tau_B F \]
respectively.

The inhomogeneous solution is given by
\[ i(t) = i_{n1} \tau_B F \]
whereas collector current will not change until the extra electrons injected at the emitter side of the base reach the collector. A more complete analysis that takes account of the distributed nature of base charging predicts that \( i_c \) does not change at all for times close to \( t = 0 \); initially, the increments in the base and emitter currents are equal. The collector current first begins to change when \( t \) reaches the base transit time \( \tau_B \). The behavior of \( i_c \) then rapidly approaches the transient solution predicted by the charge-control model and essentially matches that result for times larger than \( t = \tau_c \). (See inset in Figure 7.17.)

The foregoing example may seem artificial because of the number of simplifications that have been imposed, first in the choice of circuit and second in the approximations that have been made. These physical simplifications, however, have avoided mathematical complications that might tend to obscure the use of the model.

One such complication would arise if, for example, the collector in Figure 7.16a were not connected to an ac ground, but rather was connected to the source through a load resistor \( R_f \). Since \( V_{CB} \) would be variable in this case, the current required to charge \( Q_{VC} \) would not be negligible. The solution for \( i_c \) can be obtained readily if we define an effective capacitance \( C_E^* \) equal to the average of \( dQ_{VC}/dV_{CB} \) over the collector voltage interval. The solution for \( i_c \) is then found to be equal to Equation 7.4.9 provided that the time constant \( \tau_{BF} \) in Equation 7.4.9 is replaced by [Problem 7.19]:

\[ \tau_{BF}' = \tau_{BF} \left( 1 + \frac{R_i C_E}{\tau_f} \right) \]

To derive Equation 7.4.10, it is necessary to note that the presence of the collector capacitance affects the base current much more than it does the collector current.
is also induced. For example, under very high forward bias, the number of holes can rise above the value of the acceptor doping in the base. This condition is termed high-level injection. Equation 8 demonstrates that there is a drop in collector current. If the injected hole concentration is very much greater than the background acceptor density, then quasineutrality will be established through the relation
\[ p = n = n_t \exp(qV_{BE}/2kT) \]. This leads to an expression for \( J_c \) with voltage dependence \( \exp(qV_{BE}/2kT) \).

If the reverse-bias voltage \( V_{CB} \) is increased, then the depletion region in the base will grow, and the integrated hole density will decrease. From Eq. 8, this must lead to an increase in collector current. The change is found, for the case of uniformly doped base, to be
\[ dJ_c/dV_{CB} = J_c C_{BC}/p_b w \] (9)

\( C_{BC} \) is the base–collector depletion capacitance per unit area. The limiting expression for \( J_c \) as \( V_{CB} \) is varied, is given by
\[ J_c(V_{CB})/J_c(0) = 1/(1 - V_{CB}/V_A) \approx 1 + (V_{CB}/V_A) \] (10)

\( V_A = C_{BC}/(q/p_b w) \) is known as the Early voltage. The associated cut conductance degrades transistor voltage gain, and must be minimized maintaining a suitably large value of \( p_b w \).

The above expressions for \( J_c \) correspond to cases in which electron flow miniated by base transport. This covers most situations encountered in most of the devices. As described below, additional contributions must be added to the current model bipolar transistors where potential barriers intercept current flow, or produce injected electron distributions far from equilibrium. In the most aggressively scaled devices, further corrections are added, inasmuch as Eq. 7 is not valid if carriers experience fewer scattering as they traverse the base.

**Current.** The hole current which must be supplied from the base to fill the holes has a number of different components, corresponding to recombination of holes with electrons in different regions of the device. As pictured in Fig. 4, there is recombination in the quasineutral base, in the emitter–base depletion region, at the emitter periphery, in the emitter body, and at the emitter surface. These key components of the base current are evaluated following.

**Recombination.** Recombination of electrons and holes within the base proceed by direct photon-induced recombination, via deep levels, or by the Auger effect. Under most circumstances, a recombination rate \( \tau_{rec} \) can be determined such that the net number of recombinations per unit volume in the base is
\[ U = (n - n_{eq})/\tau_{rec}, \] where \( n \) is the density of electrons injected into the base, and \( n_{eq} \) represents the thermal equilibrium minority carrier concentration. The associated base current is the integral of the recombination density over the base:
\[ J_{b1} = qN_s/\tau_{rec} \] (11)

where \( N_s \) is the integrated density of excess electrons injected into the base. For the simple transistor with uniform base, \( N_s \) can be directly evaluated as
\[ N_s = [n(0) + n(w) - 2n_{eq}]w/2 = n_t^r [\exp(qV_{BE}/kT) + \exp(qV_{BC}/kT) - 2]w/2p_b \] (12)

In normal operation \( V_{BE} >> kT/q, V_{BC} << -kT/q \), only the term involving \( V_{BE} \) is important. \( N_s \) can be easily related to the collector current \( J_c \) for a transistor with uniform doping and composition from Eq. 1 as
\[ N_s = J_c w^2/(2D_n q) = J_c \tau_b/q \] (13)
\[ \tau_b = w^2/(2D_n) \]

Here \( \tau_b \) is the base transit time for electrons. Thus the base current \( J_{b1} \) and the associated current gain \( \beta_1 \) limited by this mechanism alone are
\[ J_{b1} = (\tau_b/\tau_{rec})J_c \]
\[ \beta_1 = J_c/1J_{b1} = \tau_{rec}/\tau_b \] (14)

The ratio of base transit time \( \tau_b \) to recombination time \( \tau_{rec} \) can be viewed as the probability of recombination in the base of a given injected electron. In more complex structures, \( N_s \) can also be found to be proportional to \( J_c \), and Eqs. 14–15 continue to hold with an appropriately modified expression for \( \tau_b \). In most modern devices, \( \tau_b \) is of order of picoseconds, whereas \( \tau_{rec} \)
is of order nanoseconds or microseconds; thus \( J_{b1} \) is generally quite small.

**Emitter–Base Depletion Region Recombination.** According to the statistics for recombination via deep levels (quantified by Shockley–Read–Hall and others\(^1\)), the effective recombination lifetime for carriers varies strongly with carrier density. The recombination rate reaches a strong maximum under the condition \( p = n \), inasmuch as both carriers are relatively plentiful for capture. Within the bipolar transistor, this occurs over a relatively thin region within the emitter–base depletion region, for which \( n = p = n_i \exp(qV_{BE}/2kT) \). Integrating the recombination density over the depletion region leads to the net base-current contribution:

\[
J_{b2} = q \int U(x)dx = (q n_i / \tau_{eh}) (2 \pi kT / q \epsilon_p) \exp(qV_{BE}/kT)
\]

(15)

where \( \epsilon_p \) is the electric field at the plane of maximum recombination. This contribution to base current thus has a voltage dependence of \( \exp(qV_{BE}/2kT) \), that is, an ideality factor of 2. Correspondingly, the current gain \( \beta_2 \) limited by this mechanism increases with increasing collector current:

\[
\beta_2 = J_c / J_{b2} \propto \exp(qV_{BE}/2kT) \propto J_c^{1/2}
\]

(16)

**Reverse Injection into Emitter.** When the base–emitter junction is forward-biased, holes flow by diffusion (and potentially drift) into the emitter, recombining with electrons in the emitter body and at the emitter surface. The carrier concentration profile differs significantly among devices, according to the thickness of the emitter relative to the hole diffusion length, and according to the recombination velocity at the surface of the emitter. The carrier profile shown in Fig. 3 corresponds to a situation with a thick emitter, where the diffusion length of holes governs their profile. More frequently, bipolar transistors have thin emitters with metallic contacts (causing high surface recombination velocity), for which the density of holes varies approximately linearly with distance within the emitter. In either case, the hole current density can be calculated taking into account diffusion, drift associated with built-in fields, and varying bandgap, with a method directly analogous to the treatment of electron flow (Eqs. 4–8). This results in hole current \( J_{b3} \) given by

\[
J_{b3} = q \int (n / n_i^2) \exp(qV_{BE}/kT) dx
\]

(17)

The DC current gain of the transistor, \( \beta_3 \), as limited by this mechanism, is

\[
\beta_3 = J_c / J_{b3} = (D_n / D_p) (\int (n / n_i^2) dx / \int (p / n_i^2) dx)
\]

(18)
where $n_e$ and $n_b$ are the intrinsic carrier concentrations in emitter and base, respectively. For uniform doping of the emitter and base, and a thin emitter with infinite surface recombination velocity, this expression yields

\[
\beta_3 = \frac{(D_e/D_p)(n_e w_e \exp(E_g/kT)/p_i w \exp(E_p/kT))}{(15)}
\]

This equation reveals a number of important factors for transistor design. Current gain is directly proportional to the ratio $n_e/p_i$, (emitter to base doping concentration). To maintain suitably high values of current gain, this ratio is typically chosen to be on the order of 1000. For a silicon homojunction transistor, emitter doping levels of $10^{19}/cm^3$ and base doping levels on the order of $10^{17}$-$10^{18}/cm^3$ are typically used. Also seen in Eq. 19 is the fact that the difference in bandgap energy between emitter and base has a major impact on current gain. In homojunction transistors, the principal bandgap difference stems from the bandgap shrinkage resulting from heavy doping on the emitter side of the device. For high doping levels, the effective bandgap energy is reduced because of the effects on carrier energy of the potentials of the donor or acceptor atoms and their associated fluctuations (bandtailing effect), and because of the binding energy of the associated electron and hole gases. The extent of bandgap shrinkage in silicon has been extensively studied theoretically and experimentally.\(^\text{15}\) As a result of the reduction in emitter bandgap, the current gain is reduced according to the factor $\exp(-\Delta E_g/kT)$, where $\Delta E_g$ is the bandgap narrowing at the emitter-base junction edge.

The overall base current is the sum of these components, plus, potentially, the additional currents associated with tunneling at the emitter–base junction and emitter–edge currents. Base current components are often difficult to distinguish in specific circumstances, although in general they differ in their dependence on $V_{BB}$, on temperature, on base thickness, and on the ratio of emitter periphery to emitter area.

1.2.2 Charge Storage

The AC and transient characteristics of bipolar transistors are dominated by the charge stored within the device that must be increased or decreased when the bias conditions are changed. Within the charge control model, the base current consists of a DC component that is associated with the instantaneous bias voltages at the device terminals and a transient value given by the derivative of the charge stored in the device. For normal device operation, this corresponds to

\[
I_0(t) = I_{00}[V(t), V_{BB}(t)] + (dQ/dt)
\]

\[
I_{i}(t) = I_{i0}[V(t), V_{BB}(t)]
\]

\[
I_{i}(t) = I_{i0}(t) + I_{i}(t)
\]
Fig. 5 Excess hole charge and excess electron charge stored in a transistor in forward bias.

Moreover, in the charge control approximation, the value of $Q_B$ is taken to be the value that would apply under steady-state conditions, given the instantaneous values of the terminal voltages. In the spirit of this approximation, from knowledge of the charge distribution in the transistor under steady-state conditions and Eq. 20, the transient and AC performance of the transistor can be calculated.

The charge stored in the transistor, $\Delta Q_B$, in excess of the charge contained in the device at zero bias, $Q_{B0}$, is pictured in Fig. 5. Here the electron and hole distributions are shown for conditions of zero bias and a representative active bias. The excess charge per unit area $\Delta Q_B$ can be computed by integrating over the entire device the contributions corresponding to excess electron charge density, or the contributions corresponding to excess hole charge. The two resulting values are equal in magnitude, since the overall transistor is neutral.

**Charge Contributions.** Distinct contributions to the charge, $Q_b$, may be associated with the different regions of the device (termed $Q_{bc1}$, $Q_{bc2}$, $Q_h$, $Q_{bc1}$, and $Q_{bc2}$ in Fig. 5). Each of these contributions is approximately proportional to collector-current density $J_c$. It is of interest to compute the ratio $Q/J_c$, a quantity with the units of time, which corresponds loosely to the transit time of carriers through the different regions of the device. These are approximately evaluated in the following.

**Emitter Region, $Q_{bc1}$**. Under forward-bias of the base-emitter junction,
EXCESS HOLES ARE INJECTED INTO THE EMITTER WITH A DISTRIBUTION THAT DEPENDS ON THE DETAILS OF THE EMITTER THICKNESS AND SURFACE RECOMBINATION (AS DESCRIBED ABOVE). FOR THE SIMPLE CASE OF THIN EMITTER WITH METAL CONTACT, THE CHARGE STORED IS

\[ Q_{be1} = \int q(p - p_{eq})dx = qw_c n_{c0}^2 \exp(qV_{BE}/kT)/(2n_c) \]

\[ \tau_{be1} = Q_{be1}/J_c = w_c w_{pe} n_{c0}^2 (2D_n n_c n_{h0}^2) \]

(21)

\( \tau_{be1} \) IS A SIGNIFICANT CONTRIBUTION TO THE OVERALL DELAY OF SILICON BIPOLAR TRANSISTORS. TO MINIMIZE IT, THIN EMITTERS WITH HIGH DOPING ARE DESIRED, WHILE LIGHTLY DOPED, THIN BASES ARE ALSO IMPORTANT (AS ALSO REQUIRED TO ESTABLISH HIGH DC CURRENT GAIN). ANOTHER EFFECTIVE STRATEGY IS TO USE A WIDE BANDBAND EMITTER, AS DESCRIBED LATER FOR HETEROJUNCTION BIPOLAR TRANSISTORS (HBTs).

**Emitter-Base Depletion Region, \( Q_{be2} \).** CHARGES MUST BE STORED AT THE EDGES OF THE DEPLETION REGION TO SUPPORT THE ELECTROSTATIC FIELDS AS THE JUNCTION VOLTAGE IS CHANGED. FOR A CHANGE IN VOLTAGE \( dV_{BE} \), THE CORRESPONDING CHANGE IN STORED CHARGE IS \( C_{BE} dV_{BE} \). \( C_{BE} \) MAY BE APPROXIMATELY EVALUATED AS THE CAPACITANCE OF A P-N DEPLETION REGION. THE DELAY TIME \( \tau_{be2} \) ASSOCIATED WITH THIS CHARGE IS

\[ \tau_{be2} = dQ_{be2}/dJ_c = C_{BE}(dU_c/dV_{BE}) = C_{BE} qJ_c/kT \]

(22)

\( \tau_{be1} \) AND \( \tau_{be2} \) ARE FREQUENTLY LUMPED TOGETHER IN A CONTRIBUTION \( \tau_c \) KNOWN AS THE EMMITER DELAY.

**Base Region, \( Q_b \).** CHARGE ASSOCIATED WITH THE ELECTRONS INJECTED INTO THE BASE IS NEUTRALIZED WITH ADDITIONAL HOLES ADDED TO THE BASE. THE OVERALL AMOUNT OF CHARGE, AND THE ASSOCIATED DELAY TIME, HAVE ALREADY BEEN DISCUSSED FOR THE CALCULATION OF BASE CURRENT. FOR A UNIFORM BASE, THEY ARE GIVEN BY

\[ Q_b = \int q(n - n_{eq})dx = J_c w^2/(2D_n) \]

\[ \tau_b = w^2/(2D_n) \]

(23)

IN MODERN DEVICES, THERE ARE A VARIETY OF CORRECTIONS THAT MUST BE CONSIDERED. WHEN SIGNIFICANT ELECTROSTATIC FIELDS ARE PRESENT IN THE BASE, ELECTRON FLOW PROCEEDS BY DRIFT AS WELL AS BY DIFFUSION, AND \( \tau_b \) IS CUSTOMARILY REPRESENTED AS \( \tau_b = w^2/\eta D_n \), WHERE \( \eta \) IS A CORRECTION FACTOR THAT DEPENDS ON THE MAGNITUDE OF THE ELECTRIC FIELD PRESENT (WITH \( \eta = 2(1 + (q' w/2kT)^{1/2}) \)).

THE VELOCITY WITH WHICH ELECTRONS MAY EXIT FROM THE BASE IS LIMITED TO A VALUE OF THE ORDER OF THE SATURATION VELOCITY, \( v_{sat} \). IN ADDITION, IN THIN BASES, DIFFUSION FLOW IS NOT GOVERNED BY THE SIMPLE FICK'S LAW EXPRESSION; RATHER, IT IS LIMITED TO A VELOCITY KNOWN AS DIFFUSION VELOCITY OR THERMIONIC-EMISSION VELOCITY. THIS VELOCITY CORRESPONDS TO THE SITUATION IN WHICH AN ENTIRE
thermal electron population is directed from emitter to collector, and there is no backscattered or returning carrier flow. Under such circumstances, an approximate expression for base transit time is given by

\[ \tau_n = \frac{1}{2D_e} + \frac{1}{v_m} \]

where \( v_m \) is the velocity at which the electrons exit the base at the collector edge. \( v_m \) is typically given by the effusion or thermionic-emission velocity of electrons, 

\[ v_m = \frac{kT}{2\pi m^*} \]

**Collector Region, \( Q_{bc1} + Q_{bc2} \).** As bias conditions are changed, the charge stored in the base-collector depletion region must change, through two mechanisms. In the first mechanism, if \( V_{bc} \) changes, then a depletion charge \( Q_{bc1} = C_{bc} \frac{dV_{bc}}{dU_c} \) must be added (or removed) to the base and to the collector edges of the depletion region. Correspondingly, there is delay time \( \tau_{bc1} \) defined as \( dQ_{bc1}/dU_c \). By convention, the charge is computed with the collector terminal incrementally short-circuited to the emitter terminal. As a result, by taking into account the external circuit, which may have series resistances, as \( V_{BE} \) is changed, the variation of \( V_{bc} \) is given by

\[ \Delta V_{bc} = \Delta V_{be} + I_c(R_E + R_C) \]  

where \( R_E \) and \( R_C \) are extrinsic parasitic resistances associated with these terminals. The collector current \( I_c \) is given by \( I_c \) times the area of the emitter, \( A_E \). The delay time is thus

\[ \tau_{bc1} = C_{bc} \frac{dV_{bc}}{dU_c} + R_E A_E + R_C A_E \]

A second mechanism causing a variation of the charge in the base (at constant \( V_{bc} \)) is associated with the finite velocity of electrons within the depletion region. As \( J_c \) increases, the electron density within the collector also increases, and the associated electron charge modifies the space-charge density distributed throughout the collector, from \( N_D \) to \( N_{Deff} = N_D - J_c/q v_s \)

(\( N_D \) is the donor concentration in the collector region, and \( v_s \) is the electron velocity, typically at its saturated value). The injected electrons act as acceptor dopants in the depletion region. The resultant change in "doping" changes the amount of charge at the base edge of the depletion region. Within the simple approximation for a one-sided \( p-n \) junction, the charge \( Q_c \) in the depletion region, and the associated time constant \( \tau_{bc2} \) are

\[ Q_c = [2e q N_{D} (V_{CB} + V_{be})]^{1/2} = [2e q (N_D - J_c q v_s)(V_{CB} + V_{be})]^{1/2} \]

\[ \tau_{bc2} = \frac{dQ_c}{dU_c} = \frac{Q_c}{(2N_{D}q v_s)} = \frac{w_c}{2v_s} \]

where \( w_c \) is the thickness of the collector depletion region. The result for \( \tau_{bc2} \) corresponds to one-half the time expected for an electron to traverse the collector depletion region, traveling at the saturated drift velocity. The factor of two accounts for the fact that the charge associated with the electrons...
distributed through the depletion region is terminated partly at the base depletion region edge, and partly at the collector depletion region edge, under conditions of constant voltage drop across the depletion region. For situations in which the velocity of the electrons varies spatially within the collector, the above simple picture may be modified to give

$$\tau_{bc} = \int [(1 - v/v_0) / v(v)] dv$$

(27)

The overall charge stored in the base is the sum of the contributions defined above. In similar fashion, the overall delay, termed \(\tau_{ee}\), the emitter-to-collector delay, is the sum of the contributions described:

$$Q_B = Q_{bc1} + Q_{bc2} + Q_B + Q_{be1} + Q_{be2}$$

$$\tau_{ee} = \tau_{be1} + \tau_{be2} + \tau_b + \tau_{be1} + \tau_{bc2}$$

(28)

**Modes of Transistor Operation.** The preceding discussion centered on bias conditions in which the emitter–base junction is forward-biased and the base–collector junction is reverse-biased or only moderately forward-biased (called the forward active mode of operation). If the base–emitter junction and base–collector junctions are both reverse-biased, the transistor is in cutoff mode, and no current flows. With the base–collector forward-biased and the base–emitter reverse biased, inverse operation of the transistor is obtained, and emitter current will flow. Under most circumstances, the doping and area ratios employed in the transistor fabrication will lead to a very low value of current gain in this mode of operation, potentially below unity. Finally, if the base–emitter junction and base–collector junctions are both forward-biased, the transistor is said to be in saturation. Representative minority-carrier concentrations in the saturation mode are shown in Fig. 6. The charge stored within the base is considerably greater than what would be stored under normal operation. Also, large amounts of minority-carrier charge are stored in the collector region. The excess charges (holes) must be disposed of during switching operations either by recombination or by extracting them from the base terminal. Typically, they slow the transistor performance dramatically. In saturation, the charge stored is partitioned into contributions (\(Q_B\) and \(Q_R\)) that are associated with emitter and collector. The base charge of these two portions can be determined from the graphical analysis of Fig. 6. The overall form of the charge control equations, including saturation, then becomes:

$$I_B(t) = I_{B0}[V_{BE}(t), V_{BC}(t)] + (dQ_t/dt) + dQ_R/dt$$

(29a)

$$I_C(t) = I_{C0}[V_{BE}(t), V_{BC}(t)] + (dQ_R/dt)$$

(29b)

$$I_E(t) = I_{E0}[V_{BE}(t), V_{BC}(t)] + (dQ_t/dt)$$

(29c)
1.2.3 Figures-of-Merit for Transistor Performance

Current Gain Cutoff Frequency \( f_T \). \( f_T \) is the frequency at which the magnitude of the transistor incremental short-circuit current gain, \( h_{fe} \), drops to unity. It is a key estimator of transistor high-speed performance. If we consider AC transistor behavior, assuming small-signal excitation voltages and currents with time dependence \( \exp(j\omega t) \), then within the charge control framework, the incremental current gain \( h_{fe} \) can be calculated as follows:

\[
\begin{align*}
  i_b &= i_{b0} + j\omega q_b \\
  i_c &= i_{c0}
\end{align*}
\]

\[
|h_{fe}| = |i_c/i_b| = |i_{c0}/(i_{b0} + j\omega q_b)| = 1/\left| (i_{b0}/i_{c0}) + (j\omega q_b/i_{c0}) \right| = 1/(1/\beta + j\omega\tau_{ce})
\]

Here we have introduced notation \( i_b, i_c, \) etc. to denote small-signal values of the quantities \( I_B, I_C, \) etc., and have noted \( q_b/i_{c0} = dQ_{B}/dI_C = \tau_{ce} \). The current gain \( h_{fe} \) has the frequency dependence noted in Fig. 7, reaching a low-frequency value of \( \beta \), and dropping at a rate of 6 dB/octave \((1/f)\) dependence) at frequencies above the beta-rolloff frequency \( 1/(2\pi\beta\tau_{ce}) \). The current gain magnitude drops to unity at a frequency \( f_T = 1/(2\pi\tau_{ce}) \), an expression arrived at by neglecting \( 1/\beta \) compared with unity. The delays studied in Section 1.2.2 thus have a key role in governing the AC current gain of the device. A brief summary of the contributions to \( f_T \) is frequently written, for normal operation at moderate current densities, as

\[
1/(2\pi f_T) = \tau_f + \tau_{c}' + \left( \tau_{s}/2\tau_{s} \right) + (R_E + R_C)C_{BC} + \left[ (C_{BE} + C_{BC})kT/(qI_c) \right]
\]

The first three contributions are sometimes referred to as \( T_F \) in circuit level models, as described in the following. The last component depends on collector \( c \) at low cur experimer.

Fig. 7  Re gain, \( h_{fe} \).
collector current as $1/I_C$, and leads to a dramatic slowdown of the transistors at low current levels. To distinguish the contributions, it is customary to plot experimental data of $f_T$ vs $I_C$ in the form of $1/(2\pi f_T)$ vs $1/I_C$, as shown in Fig. 8; the slope can be identified with $kT\ln(C_{H1} + C_{Hc})$. As the current $I_C$ reaches very high levels, $f_T$ drops ($\tau_{ce}$ rises) due to the base pushout effect described below.

**Maximum Frequency of Oscillation ($f_{max}$).** This is the frequency at which the maximum available power gain of the transistor drops to unity. $f_{max}$ is widely useful to estimate power gain, since over a wide range of frequencies, maximum available power gain, $G_p$, follows the relation

$$G_p = (f_{max}/f)^2$$

(32)

$f_{max}$ is different from (and typically larger than) $f_T$, because in addition to current gain, $f_{max}$ takes into account the possibility of voltage gain. A simple analysis based on the simplified hybrid pi equivalent circuit for the bipolar transistor\(^1\) provides the basis for estimating the factors important to $f_{max}$. The
\[ J = \text{constant} \]

Base \hspace{2cm} Collector

\[ v(\ast) \]

\[ n(x) \]

\[ \Delta n(x) \]

\[ x = 0 \]

\[ x \]

\[ x = w \]
The problem is to find $Q_{bc_2}$, the variation in the base charge (associated with the base-collector depletion region).

Approach: Find the induced charge because of the sheet $\Delta n(x)$ in the width $\Delta x$ and integrate from $x = 0$ to $x = w$.

The voltage across the $B-C$ junction is $(V_{cb} + V_{bj})$, constant (not perturbed by the charge).

But the charge perturbs the electric field such that

$$
\Delta E^+ - \Delta E^- = \frac{q\Delta n(x) \Delta(x)}{\varepsilon}
$$

(Gauss’ Law)

Also, $\Delta E^+ \cdot x + \Delta E^- (w - x) = 0$ (no change in voltage)
\[ \therefore \Delta E = -\frac{\Delta E^+}{w-x} \]

\[ \therefore \Delta E^+ - \left( -\Delta E^+, \frac{x}{w-x} \right) = \frac{q\Delta n(x)\Delta x}{\varepsilon} \]

But \( \Delta n(x) = \frac{J_c}{qV(x)} \)

\[ \therefore \Delta E^+ \left( 1 + \frac{x}{w_c - x} \right) = \frac{J_c}{V(x)} \cdot \frac{\Delta x}{\varepsilon} \]

or \( \Delta E^+ = \Delta \frac{Q_{bc}}{\varepsilon} \cdot J_c \cdot \left( \frac{w_c - x}{w_c} \right) \frac{dx}{V(x)} \)

\[ \Delta Q_{bc} = J_c \int \left( 1 - \frac{x}{w_c} \right) \frac{1}{V(x)} dx \]

\[ \Delta Q_{bc} = J_c \int_0^1 \left( 1 - \frac{x}{w_c} \right) \frac{1}{V(x)} dx \]

\[ \frac{dQ_{bc}}{dJ_c} = \tau_{bc} = \int_0^1 \left( 1 - \frac{x}{w_c} \right) \frac{1}{V(x)} dx \]
Current Gain Cutoff Frequency \( (f_T) \). \( f_T \) is the frequency at which the magnitude of the transistor incremental short-circuit current gain, \( h_{fe} \), drops to unity. It is a key estimator of transistor high-speed performance. If we consider AC transistor behavior, assuming small-signal excitation voltages and currents with time dependence \( \exp(j\omega t) \), then within the charge control framework, the incremental current gain \( h_{fe} \) can be calculated as follows:

\[
\begin{align*}
    i_b &= i_{b0} + j\omega q_b \\
    i_c &= i_{c0} + j\omega q_c
\end{align*}
\]

Here we have introduced notation \( i_b \), \( i_c \), etc. to denote small-signal values of the quantities \( I_b, I_c \), etc., and have noted \( q_b \), \( q_c \) etc. The current gain \( h_{fe} \) has the frequency dependence noted in Fig. 7, reaching a low-frequency value of \( \beta \), and dropping at a rate of 6 dB/octave \( (1/f) \) dependence at frequencies above the beta-rolloff frequency \( 1/f \). The current gain magnitude drops to unity at a frequency \( f_T = 1/(2\pi\tau_{\omega}) \), an expression arrived at by neglecting \( 1/\beta \) compared with unity. The delays studied in Section 1.2.2 thus have a key role in governing the AC current gain of the device. A brief summary of the contributions to \( f_T \) is frequently written, for normal operation at moderate current densities, as

\[
\frac{1}{2\pi f_T} = \tau_{w} + \tau_{c} + \left(\frac{w}{2v_{c}}\right) + \left(R_{e} + R_{c}\right) C_{ac} + \left[\left(C_{ae} + C_{ac}\right) kT/q\right]
\]

![Graph](image.png)
collect current as $1/I_c$, and leads to a dramatic slowdown of the transistors at low current levels. To distinguish the contributions, it is customary to plot experimental data of $f_r$ vs $I_c$ in the form of $1/(2\pi f_r)$ vs $1/I_c$, as shown in Fig. 8; the slope can be identified with $kT/q (C_{BE} + C_{BC})$. As the current $I_c$ reaches very high levels, $f_r$ drops ($\tau_e$ rises) due to the base pushout effect described below.

**Maximum Frequency of Oscillation ($f_{\text{max}}$).** This is the frequency at which the maximum available power gain of the transistor drops to unity. $f_{\text{max}}$ is widely useful to estimate power gain, since over a wide range of frequencies, maximum available power gain, $G_p$, follows the relation

$$G_p = (f_{\text{max}}/f)^2$$

(32)

$f_{\text{max}}$ is different from (and typically larger than) $f_r$, because in addition to current gain, $f_{\text{max}}$ takes into account the possibility of voltage gain. A simple analysis based on the simplified hybrid pi equivalent circuit for the bipolar transistor provides the basis for estimating the factors important to $f_{\text{max}}$.

**Delays in Bipolar transistors**

1. First Consider diodes
   A. Long Base diode (has only capacitive delay)

\[ C = \frac{dQ}{dV} \quad \text{space charge Capacitance} = \frac{\varepsilon}{w(v)} \]
\[ x = 0 \]

\[ Q = qA \int_0^x \Delta e(x') dx' = qA \int_0^x \Delta p(0) e^{-\int \mu_d dx'} \]

\[ Q = qA \Delta p(0) L_p \]

\[ I_p(x' = 0) = qA \frac{\Delta p(0)}{L_p} \]

\[ Q = \frac{I_p(0) L_p^2}{\Delta p} = I_p \tau_p \]

Current in charge analysis is always considered to be the rate of which carriers have to be supplied to maintain in the charge distribution.

\[ \frac{dQ_p}{dV} = \frac{dI_p}{dV} \tau_p = \frac{\tau_p}{\tau_e} \]

Time dependent continuity example:

\[ qA \frac{\partial (\Delta p)}{\partial t} = -qA \frac{\Delta p}{\tau_p} \frac{\partial I_p}{\partial x} \]

\[ \frac{\partial (\Delta p)}{\partial t} = -\frac{\Delta p}{\tau_p} + D_p \frac{d^2 \Delta p}{dx^2} \quad \text{(Assumes only diffusion)} \]

Assume a solution \( \Delta p = \Delta p(0) e^{-\int \mu_d dx'} + \Delta p e^{j \omega w} \)

\[ j \omega \Delta p = -\frac{\Delta p}{\tau_p} + \Delta p \frac{d^2 \Delta p}{dx^2} \]

solution. \( \Delta p = C_1 e^{-\int \mu_d dx'} + C_2 e^{j \omega w} \)

\[ \lambda = \left( \frac{j \omega}{\Delta p} + \frac{1}{\Delta p \sigma_p} \right)^{1/2} \quad \text{(Frequency dependent diffusion length)} \]

\( C_2 = 0 \) for physical reasons
\[ C_i = \Delta \rho (0) \]
\[ \tilde{\Delta} \rho = \Delta \rho (0)e^{-\frac{q\omega \tau_p}{\Delta \rho \tau_p}} \]
\[ \frac{\mathcal{I}}{x = 0} = -qA\Delta \rho \frac{d\Delta \rho}{dx} \bigg|_{x = 0} = qA\Delta \rho (0)\sqrt{\frac{j\alpha}{\Delta \rho} + \frac{1}{\Delta \rho \tau_p}} = \frac{qA\Delta \rho \tilde{\Delta} \rho (0)}{2} \]

Small-signal approximation: \( V_s = V_{dc} + \tilde{\mu} \)
\[ \Delta \rho (0) = \Delta \rho (0) + \tilde{\Delta} \rho (0) = p_n e^{\frac{q(V_{dc} + \tilde{\mu})}{kT}} = p_n e^{\frac{q\tilde{\mu}}{kT}} \left[ e^{\frac{q\tilde{\mu}}{kT}} \right] \]

Since we are assuming \( \tilde{\mu} \) to be small, \( e^{\frac{q\tilde{\mu}}{kT}} = 1 + \frac{q\tilde{\mu}}{kT} \)
\[ \Delta \rho (0) = \Delta \rho (0) \frac{q\tilde{\mu}}{kT} \]

We know \( I_d = \frac{qA\Delta \rho \Delta \rho (0)}{L_p} \Rightarrow \tilde{\Delta} \rho (0) = \frac{L_pI_{dc} \frac{q\tilde{\mu}}{kT}}{qA\Delta \rho \Delta \rho (0)} \]
\[ y = \frac{\mathcal{I}}{\tilde{\mu}} = \frac{qI_{dc}}{kT} \left[ 1 + \frac{j\omega \tau_p}{r_e} \right] = \frac{1}{r_e} \frac{1}{1 + \frac{j\omega \tau_p}{r_e}} \]

For sufficiently small frequencies,
\[ y = \frac{1}{r_e} \left[ 1 + \frac{j\omega \tau_p}{2} \right] = G + j\omega C \]
\[ C = \frac{\tau_p}{2r_e} \quad \text{This is the diffusion capacitance.} \]

We have ignored the depletion region capacitance.
\[ y_{\text{eff}} = \frac{1}{r_e} + j\omega \left( C_{\text{diff}} + C_{\text{dep}} \right) \]
\[ y = \frac{1}{r_e} + j\omega \left( \frac{\tau_p}{2r_e} + \frac{\varepsilon A}{\omega} \right) \]

Let's look again at p-n junctions

**p-n junctions**

![Long base diode diagram](image)

Apparent capacitance  
\[ C_{\text{app}} = \frac{dG_j}{dV_n} = \frac{\tau_p}{r_e} \]

\[ C_{\text{diff}} = \frac{1}{2} \frac{\tau_p}{r_e} \]

Using \( y = G + j\omega \) allows an equivalent circuit

To get \( y \), need to solve time dependent continuity eg.

\[ \frac{dp}{dt} = \frac{1}{q} \nabla \cdot \mathbf{J} + G - R \]

\[ p = p_d + p_{ac}, \quad p_{ac} = \bar{p} = \Delta p \ e^{-\imath \omega t} \]

Small signal analysis

\[ pn = p_p e^{\imath \omega t} \approx p_p \left( \frac{\xi_p}{e^{\imath \omega t}} + e^{-\imath \omega t} \right) \]
\[ p_n \left( \frac{qV}{e \text{m}} \right) = p_n \left( 1 + \frac{qV}{kT} \right) \]

\[ y = \sqrt{1 + j\omega \tau_p \over r_c} \quad \text{at low frequency} \quad = {1 \over r_c} + j {\omega \tau_p \over 2 r_c} \]

for high frequency, \[ y = \sqrt{\omega \tau_p \over r_c} = \sqrt{\omega \tau_p \over 2 r_c} + j \sqrt{\omega \tau_p \over 2} = {1 \over r_c} \sqrt{\omega \tau_p \over 2} + j {\omega \over r_c} \sqrt{\tau_p \over 2 \omega} \]

Both small signal resistance & capacitance decrease with \( \omega \) as \( \frac{1}{\sqrt{\omega}} \).

Low frequency

High frequency

**Arbitrary Diode Length**

\[ \frac{\partial n}{\partial t} = D_n \frac{\partial^2 n}{\partial x^2} \quad \text{(Assuming negligible recombination)} \quad G - R = 0 \]

Assume, \( n(x, t) = n_s (x) + n_a (x) e^{\hbar t} \) \quad \text{(Diffusion wave)}

Considering only \( a e \),

\[ \frac{d^2 n_a}{dx^2} = \frac{n_0 \hbar}{\lambda e^2} \quad \text{where} \quad \frac{1}{\lambda e} = \sqrt{\frac{\hbar \omega}{D e}} \]

\[ n_a (x) = C e^{\pm x/\lambda e} \]

Apply Shockly B.C. \( n_a (w_{B_p}) = 0 \)
\[ n_n(x) = \frac{n_n(0) \sin h \left( \frac{w_{Bn} - x}{\lambda_e} \right)}{\sin h \left( \frac{w_{Bn}}{\lambda_e} \right)} \]

\[ J\omega(0) = q D e \frac{dn_n}{dx} \bigg|_{x=0} = -q D e \omega(0) \frac{w_{Bn}}{\lambda_e} \coth \left( \frac{w_{Bn}}{\lambda_e} \right) \]

Low frequency case:
\[ \frac{1}{\lambda_e} = \sqrt{\frac{\omega}{\Delta e}} = 1 + j \sqrt{\frac{\omega}{2De}} \]

if \( w_{Bn} \ll \Lambda e \), \[ \frac{1}{\Lambda e} = \sqrt{\frac{\omega}{2De}} \]

for \( \frac{w_{Bn}}{\Lambda e} \ll 1 \), and using \( \alpha \coth x = 1 + \frac{\alpha^2}{3} - \frac{\alpha^4}{45} + O(\alpha^5) \)

\[ J\omega(0) = -q D e n_n(0) \frac{1}{w_{Bn}} \frac{w_{Bn}}{\lambda_e} \coth \left( \frac{w_{Bn}}{\lambda_e} \right) \]

[ Injected AC current ]

\[ = \frac{-q D e n_n(0)}{w_{Bn}} \left[ 1 + \frac{j\omega w_{Bn}^3}{3De} \right] \]

Small signal approx \( n_n(0) = n_{nc} \frac{q\nu \omega}{kT} \) (Remember this)

This comes from:
\[ \nu = \nu_{DC} + \tilde{\nu} \Rightarrow n_p(0, e) = n_{pe} e^{\frac{q\nu}{kT}} = n_{pe} e^{\left( \frac{q\nu_{DC} + \nu}{kT} \right)} \]

\[ = n_{pe} e^{\frac{q\nu_{DC}}{kT}} \left( 1 + \frac{q\tilde{\nu}}{kT} \right) \]

\[ \tilde{\nu} = \nu \omega e^{\nu e} \]
\[ J \omega (0) = -[G + j \omega C] V \omega \]

\[ G = \frac{qD_{DC}}{w_{p_b}} - \frac{q}{kT}, \quad C = \frac{q^2_{DC} w_{p_b}}{3} - \frac{q}{kT} \]

But we know that

\[ \frac{dQ}{dV_{b}} = \frac{q^2_{DC} w_{p_b}}{2} - \frac{q}{kT} \]

\[ C_{apparent} = \frac{2}{3} C_{diff} \]

**Collector Current**

What is the delay in \( J_c(\omega) \) relative to \( J_e(\omega) \)?

\[ J_c(w_{p_b}) = J_{\omega}(0) \quad \text{at} \quad w_{p_b} \]

By definition

\[ J_c(w_{p_b}) = \frac{qD_{e}}{\lambda_e} n_{0_b}(0) \frac{1}{\sin h \left( \frac{w_{p_b}}{\lambda_e} \right)} \]

\[ J_c(w_{p_b}) = \frac{qD_{e}}{w_{p_b}} n_{0_b}(0) \left[ 1 - \frac{j \omega w_{p_b}^2}{6D_e} \right] \quad \text{expanded} \]

\[ J_{B_b} = \text{Base current as} \quad f(\omega) \]

\[ J_{B_b} - J_{C_b} = J_{B_b} \]

\[ J_{B_b} = \frac{qD_{e}}{w_{p_b}} n_{0_b}(0) \left[ f + \frac{j \omega w_{p_b}^2}{3D_e} - f + \frac{j \omega w_{p_b}^2}{6D_e} \right] \]
Recall

\[ J_{Bo} = -qDe \frac{n_s(0)}{w_{pp}} \left[ \frac{j\omega w_p}{2De} \right] \]

\[ J_{su} = -j\omega \frac{1}{2} qn_s \left( 0 \right) w_{pp} \]

\[ I = qA_{p}(x) v = -q \Delta p \frac{dp}{dx} \]

\[ v = \frac{I}{qA_{p}} \]

Total delay = \( \int_{0}^{v_{ps}} \frac{1}{\text{velocity}} \, dx \)

\[ \gamma_{D} = \frac{1}{I} \int_{0}^{v_{ps}} qA_{p} \, dx \]

\[ \gamma_{D} = \frac{Q_{s}}{I} \quad \text{(Transit time)} \]
\[ J_{\text{BA}} = J_{\text{EA}} - J_{\text{CB}} = j \omega \frac{1}{2} q n_0 (0) W_{p} \]

\[ J_{\text{BA}} = j \omega C_B \]

\[ J_{\text{CA}} = \left[ G - j \omega \frac{C_B}{3} \right] v_{\text{B}} \]

\[ Ac \text{ Input current} = \bar{I} = i_{\text{BA}} = g m v_{\text{B}} \]

\[ i_B \text{ supplies all stored charge variations in the base} \]

1) \( \partial Q_b \) (stored minority charge)

2) Charging \( C_{be} \)

3) Charging \( C_{bc} \)

- a) Depletion charge contributed by \( v_{bc} \)

- b) Depletion charge controlled by \( J_c \)
2.4) **High-Frequency Diffusion Currents**

2.4.1) **Diffusion Waves**

At high frequencies, the \( \partial n/\partial t \) and \( \partial p/\partial t \) terms in the continuity equations (2-21) can no longer be set to zero.

In the present Section we continue to neglect recombination/generation effects, and we continue to assume that the minority carrier flow is by pure diffusion, with zero electric field. The continuity equations in one dimension then assume the simple form

\[
\frac{\partial n}{\partial t} - \frac{\partial}{\partial x} \left( J_e - q \right) = D_e \frac{\partial^2 n}{\partial x^2} ,
\]

(10a)

\[
\frac{\partial p}{\partial t} - \frac{\partial}{\partial x} \left( J_h + q \right) = D_h \frac{\partial^2 p}{\partial x^2} ,
\]

(10b)

We look here for solutions of the form

\[
n(x,t) = n_0(x) + n_{\omega} \left(x\right) e^{j\omega t} .
\]

(11a)

\[
p(x,t) = p_0(x) + p_{\omega} \left(x\right) e^{j\omega t} .
\]

(11b)

The dc parts of these are the same as in Sec. 2.2. If we insert the ac parts of (11a) into (10a), the result may be written in the form

\[
\frac{d^2 n_\omega}{dx^2} = \frac{n_0}{\kappa_e^2} , \quad \frac{d^2 p_\omega}{dx^2} = \frac{n_{\omega}}{\kappa_e^2} ,
\]

(12)

where

\[
\frac{1}{\kappa_e} = \sqrt{\frac{\omega}{2D_e}} = (1+i) \sqrt{\frac{\omega}{2D_e}} \quad \text{and} \quad \frac{1}{\Lambda_e} = \sqrt{\frac{\omega}{2D_e}} .
\]

(13a,b)

The solutions of (12) are evidently linear superpositions of waves of the two forms
\[ n_n(x) = C \cdot \exp \left( \pm \frac{x}{\lambda_e} \right) = C \cdot \exp \left( \pm \frac{(1+i)x}{\Lambda_e} \right), \quad (14) \]

which correspond to attenuated waves towards either the left (+) or the right (-). An analogous solution is found for the hole distribution.

The simplest case is that of an infinitely long semiconductor, in which case we have (for electron diffusion into the \( p^- \) type semiconductor extending to the right)

\[ n_n(x) = n_n(0) \cdot \exp \left( \frac{x}{\lambda_e} \right) = n_n(0) \cdot \exp \left( \frac{(1+i)x}{\Lambda_e} \right). \quad (15) \]

To obtain an idea about the overall scale of the attenuation effects, assume, as an example, \( D_e = 100 \text{ cm}^2/\text{s} \), and \( f = \omega/2\pi = 3 \text{ GHz} \). This yields \( \Lambda_e \equiv 10^{-7} \text{ cm} = 1 \mu \text{m} \).

We consider here again the case of a semiconductor of finite thickness \( w_{B_e} \), at the end of which the electron concentration is pinned at its equilibrium value. Inasmuch as the equilibrium density is a steady-state density, we have the boundary condition

\[ n_n(w_{B_e}) = 0. \quad (16) \]

The appropriate solution of (12) may be written

\[ n_n(x) = n_n(0) \cdot \frac{\sin h \left[ \frac{(w_{B_e} - x)}{\lambda_e} \right]}{\sin h \left[ \frac{w_{B_e}}{\lambda_e} \right]}, \quad (17) \]

where \( n_n(0) \) is the complex amplitude of the ac part of the electron concentration at the entrance into the \( p^- \) type semiconductor. We leave its value open for now.

### 2.4.2) AC Currents

Associated with an electron distribution of the form (11a) is a current density of the form

\[ J_{cs}(x,t) = J_0(x) + J_{ac}(\xi) \cdot e^{i\omega t}. \quad (20) \]
We are interested here only in the ac part, which is given by

\[ J_{\omega} (x) = +qD_x \frac{dn_0}{dx} \]  

(21)

If we insert (17) and evaluate the derivative at \( x = 0 \), we obtain the complex amplitude of the ac part of the diffusion current entering the \( p^- \) type semiconductor.

\[ J_{\omega} (0) = -\frac{qD_x n_0 (0)}{w_{B_x}} \coth \left[ \frac{w_{B_x}}{\lambda_x} \right] \]  

(22)

Evidently, the current amplitude is complex, indicating that the current has both a conductive (real) and capacitive (imaginary) part.

We wish to determine the capacitive part to the lowest order. To this end we assume that the frequency is sufficiently low that \( \Lambda_x \gg w_{B_x} \). In that limit we may expand the hyperbolic cotangent according to its Taylor expansion, which may be written

\[ \alpha \coth \alpha = 1 + \frac{\alpha^2}{3} - \frac{\alpha^4}{45} + O(\alpha^6). \]  

(23)

If we keep only terms up to \( \alpha^2 = \left( \frac{w_{B_x}}{\Lambda_x} \right)^2 \), (22) becomes

\[ J_{\omega} (0) = -\frac{qD_x n_0 (0)}{w_{B_x}} \left[ 1 + \frac{w_{B_x}}{3D_x} + O(\alpha^2) \right]. \]  

(24)

The value of \( n_{\omega} (0) \) obviously depends on the magnitude of the ac voltage. We set

\[ V_f (t) = V_0 + V_0 e^{i\omega t}, \]  

(25)

and we assume again Shockley boundary conditions, which may be written

\[ n_p (0, t) = n_{p0} \exp \left( \frac{qV_f (t)}{kT} \right) = n_{p0} \exp \left( \frac{qV_0 (0)}{kT} \right) \exp \left( \frac{qV_0 e^{i\omega t}}{kT} \right). \]  

(26)

The second exponential may always be expanded:

\[ \exp \left( \frac{qV_0 e^{i\omega t}}{kT} \right) = 1 + \frac{qV_0}{kT} e^{i\omega t} + \left( \frac{qV_0}{kT} \right)^2 e^{2i\omega t} + O(V_0^3). \]  

(27)
Note that successive terms in this expansion correspond to terms at multiples of the fundamental frequency $\omega$. These terms give rise to additional terms in both the electron concentration and the electron current, at the various multiples of the fundamental frequency. The treatment of those higher terms is the same as that of the fundamental frequency terms, except that $\omega$ must everywhere be replaced by the appropriate multiple. We neglect those higher-frequency terms here, and write

$$\exp\left(\frac{qV_0 e^{i\omega t}}{kT}\right) = 1 + \frac{qV_0}{kT} e^{i\omega t}. \quad (28)$$

We may then write (24) as

$$n_p(0,t) = n_0 + n_0 e^{i\omega t}, \quad (29)$$

where

$$n_0 = n_{p0} \exp\left(\frac{qV_0}{kT}\right) \quad (30)$$

and

$$n_0 = n_{p0} \exp\left(\frac{qV_0}{kT}\right) \frac{qV_0}{kT} = n_0 \frac{qV_0}{kT}. \quad (31)$$

If we insert (31) into (24), we obtain

$$J_0(0) = -\frac{qD n_0}{kB} \frac{q}{kT} \left[1 + \frac{\omega^2}{3D_r} \right] V_0. \quad (32)$$

This may always be written in the form

$$J_0(0) = -(G + i\omega C) V_0. \quad (33)$$

where

$$G = \frac{qD n_0}{kB} \frac{q}{kT} \quad (34)$$

is a differential conductance and
\[ C_{diff} = \frac{q n_e W_{B_e} q}{3 kT} \]  

is a **differential conductance**, the electron part of what is commonly called the **diffusion capacitance** of the \( p - n \) junction. Analogous contributions to the differential conductance arise from the hole diffusion into the \( n - \) type side.

**Collector Currents:**

The above calculation pertained to the ac diffusion current entering the base from the emitter. At finite frequencies, this current is not equal to the current leaving the base at the collector. If we recalculate the ac diffusion current density at \( x = w \), rather than at \( x = 0 \), we obtain, instead of (9-11),

\[
J_{c_{0a}} = -J_{c_{0}}(w_{B}) = \frac{q D_{n_{0}}(0)}{h_{e}} \frac{1}{\sin h[\frac{w_{B}}{h_{e}}]}.
\]

An expansion by powers of the frequency, yields, instead of (9-13),

\[
J_{c_{0a}} = \frac{q D_{n_{0}}(0)}{w_{B}} \left[ 1 - i \omega \frac{w_{B}^{2}}{6 D_{e}} + O(\omega^{2}) \right].
\]

The difference between the two currents is of course the base current:

\[
J_{b_{0a}} = J_{E_{0}} - J_{C_{0}} = \frac{q D_{n_{0}}(0)}{w_{B}} \left[ i \omega \frac{w_{B}^{2}}{2 D_{e}} + O(\omega^{2}) \right] = i \omega - \frac{1}{2} q n_{0}(0) w_{B}.
\]

We note that the base current is purely capacitive. A little reflection shows that the capacitance is simply that capacitance that is associated with the increment in electron-hole pair storage in the base region, as a result of the increase in \( V_{bb} \) by \( V_{b} \). It is therefore called the (base) storage capacitance,

\[
C_{b} = \frac{q n_{e} W_{B_{e}} q}{2 kT}.
\]

With the help of (9-23) and (9-21), we may re-write the expression (9-26) for the collector current density as

\[
J_{c_{0a}} = \left[ G - i \omega \frac{C_{b}}{3} \right] V_{b}.
\]
Note that the capacitive part of this corresponds to a negative capacitance!

This seemingly bizarre behavior is simply a consequence of the fact that any rise in emitter current appears at the collector with a delay of one transit time. Note that the capacitance seen at the emitter terminal is always positive, namely

\[ C_{\text{diff}} = 2C_B/3. \]  \hspace{1cm} (\textbullet \text{9-29})

9.3) Equivalent Circuit Representation.

The overall behavior of the diffusion currents is easily represented by generalizing the equivalent circuit of Lecture #6, containing a current conveyor isolating the emitter from the collector. We must evidently add the base storage capacitance between the emitter and the base, and the complex transconductance in (\textbullet \text{9-28}) with its negative capacitance between the emitter and the current conveyor:

![Equivalent Circuit Diagram]

Note that this equivalent circuit describes only the diffusion currents. To this circuit we must add the space charge layer capacitances on both the emitter and the collector side, as well as the contributions from the back injection current and the base recombination currents. These additions are left as an exercise to the reader.
The central idea is the following. As the voltage is increased, the current will build up until the traveling negative space charge causes the field at the injecting contact (the emitter) to drop to zero.

In those cases that are of practical interest, the fields throughout most of the semiconductor body are large enough that the current is carried by drift. We shall therefore neglect diffusion altogether, even in the low-field region near the injecting electrode, where strictly speaking, diffusion cannot be neglected. It can be shown that this leads to only a small error in the overall current-voltage relation, but it vastly simplifies the mathematical treatment.

Our point of departure is Poisson’s Equation

\[ \frac{dE}{dx} = \rho / \varepsilon = J_v / \varepsilon v_{sat} . \quad (13-15) \]

We integrate, with the initial condition \( E(0) = 0 \), which yields

\[ E(x) = \left( \frac{J_v}{\varepsilon v_{sat}} \right) x . \quad (13-16) \]

Integrating once more yields the voltage:

\[ V = \left( \frac{J_v}{2\varepsilon v_{sat}} \right) L^2 . \quad (13-17) \]

We solve for \( J_v \) to obtain the current-voltage characteristic

\[ J_v = 2\varepsilon v_{sat} V / L^2 , \quad (13-18) \]

a linear (Ohm’s-Law type) dependence. The characteristic feature of this relation is the quadratic rather than linear dependence on length.

13.6) Velocity Saturation Limits on Transistor Speed

The Finite Collector Transit Time

The effect of the finite collector transit time is quite different from that of the finite base transit time, because inside the base-to-collector depletion layer the electrons travel in a medium where they are not neutralized by majority carriers. In effect, the electrons are traveling inside a capacitor. While an electron is in transit inside a capacitor, an electrical current is actually flowing in the leads to that capacitor.
The current is most easily calculated by not considering a single electron, but the effect of an entire sheet of charges, at a distance \( x \) from the left-hand plate, moving with the speed \( v \). (See Fig. 6)

**Fig. 6 Traveling sheet of charge**

We assume that the electron concentration inside the capacitor is a function of \( x \) only, \( n(x) \). A sheet of infinitesimal thickness \( \Delta x \) at the position \( x = X \) then carries the sheet charge density \( \Delta Q = qn(X) \Delta x \). This sheet contributes an electric field \( \Delta E_+ \) in the region \( 0 < x < X \), and a different field \( \Delta E_- \) in the region \( X < x < w \). From Gauss’ Law,

\[
\Delta E_+ - \Delta E_- = -qn(X) \Delta X / \epsilon. \tag{13-19}
\]

Also, if the voltage between the capacitor plates is held constant by the external voltage source, we have further

\[
X \cdot \Delta E_+ + (w - X) \cdot \Delta E_- = 0. \tag{13-20}
\]
As the space charge distribution travels, the total charge contained in our sheet will of course remain the same, \( d \left[ n(X) \Delta X \right]/dt = 0 \). By taking the time derivative of (13-19) and (13-20), one finds readily

\[
d \Delta E_+ /dt = d \Delta E_- /dt = v (\Delta E_+ - \Delta E_-)/w = -q \cdot v \cdot n(X) \cdot \Delta X /\varepsilon w.
\]  

(13-21)

Here \( v = dX/dt \) is the speed with which the sheet travels. Associated with this time dependence of the two \( \Delta E \) is a displacement current density

\[
\Delta J = dD/dt = \varepsilon E/dt = -q \cdot v \cdot n(X) \cdot \Delta X /w.
\]  

(13-22)

But this is of course the current per unit capacitor area flowing in the external leads of the capacitor, due to the contribution of our sheet. The total current per unit area is obtained by integration over all sheets:

\[
J = -q \cdot (v/w) \cdot \int n(X) \cdot dX.
\]  

(13-23)

**Attenuation and Phase Shift of a Traveling Electron Wave**

We apply (13-23) to a traveling electron wave of the form

\[
n(x,t) = n_0 \cdot \exp \left[ j \omega (t - x/v) \right].
\]  

(13-24)

It obviously corresponds to a wave of (angular) frequency \( \omega \) traveling with the uniform speed \( v \). The convection current in the plane \( x = 0 \) is evidently

\[
J_0 = -q n_0 v \cdot \exp (j \omega t).
\]  

(13-25)

This is the current density that would be flowing if the capacitor were infinitesimally thin and the transit time of the electrons through the capacitor were zero. With the help of (13-25) we may write (13-24) as

\[
n(x,t) = -(J_0/qv) \cdot \exp (-j \omega x /v).
\]  

(13-26)

If this is inserted into (13-23), and the integration executed, one finds readily

\[
J = J_0 \left[ \exp (-j \omega t) - 1 \right]/(-j \omega),
\]  

(13-27)

where we have introduced the electron transit time through the capacitor,
\[ \tau = \frac{w}{v}. \]  

The expression (13-27) is easily transformed into the product of an amplitude factor and a phase factor:

\[ J = J_0 \left[ \sin \left( \frac{\omega t}{2} \right) / \left( \frac{\omega t}{2} \right) \right] \exp \left( - \frac{j \omega t}{2} \right). \]  

The two factors following \( J_0 \) indicate the attenuation and the phase shift of the current leaving the capacitor by the finite transit time through the capacitor.

We note first of all that the signal delay is only one-half the transit time of the electrons themselves. We also note that there is attenuation, due to destructive interference between different portions of the traveling wave. For \( \omega = \frac{2\pi}{\tau} \) the amplitude factor is zero, and no current is collected at all. This is the case when the wavelength \( \lambda = 2\pi v/\omega \) of the traveling wave is equal to the capacitor plate separation \( w \).

Clearly, a transistor is restricted to frequencies significantly below \( \omega = \frac{2\pi}{\tau} \) for useful operation. However, we shall see later that one device, the Impact Avalanche Transit Time (IMPATT) diode actually utilizes the strong phase shift just below that frequency.

For sufficiently small frequencies we may, as usually, expand our result by powers of \( j \omega t \). Carrying terms up to order \( (j \omega t)^2 \), we obtain

\[ \frac{J}{J_0} = 1 - j \omega t/2 - (j \omega t)^2 / 6 = 1 - j \omega t/2; \]  

\[ \left| \frac{J}{J_0} \right|^2 = 1 - 2(j \omega t)^2 / 6 + (j \omega t)^2 / 4 = 1 - (j \omega t)^2 / 2; \]  

\[ \left| \frac{J}{J_0} \right| = 1 - (j \omega t)^2 / 24. \]