Natural
Ge
Sc
GaAs

Eq (T=300 K)
0.803
1.12
1.43

Breakdown field (kV/cm)
$\approx 10^5$
$\approx 3 \times 10^5$
$\approx 4 \times 10^5$

Consider the effect of ionization on an IMPATT.

\[
\text{Assume } \alpha = \beta = \alpha
\]

Small reverse electron saturation current flows under the quiescent reverse bias.

\[n_e = \frac{1}{\alpha} \frac{dn}{dx} + \text{generation} \]

Electron continuity equation

\[# \text{electrons generated per second by acceleration} \]

\[G_n(e) = \alpha \cdot n \cdot v \]

# of electrons generated per electron per cm$^{-1}$

# of available electrons (# of carriers handled)
The equations represent the following:

\[ G_{in}(h) = \alpha (n \nu_{S} + p \nu_{S}) \]

\[ \frac{\partial n}{\partial t} = \frac{1}{q} \frac{\partial}{\partial x} \left( \nu \frac{\partial n}{\partial x} \right) \]

Assume \( \frac{\partial}{\partial t} = 0 \) and assume diffusion \( \beta = 0 \).

\[ J_{\text{total}} = J_{n} + J_{p} \]

Under steady state conditions,

\[ 0 = \frac{\partial J_{n}}{\partial x} + \alpha (q_{n} \nu_{S} + q_{p} \nu_{S}) \]
We know that \( J_n = q'v_n v_s \) and \( J_{Total} = J_n + J_p \)

\[
\Rightarrow \frac{dJ_n}{dx} = x J_{Total}
\]

The extent \( D \) of the avalanche region is \( W_a \)

\[
\int_0^{W_a} dJ_n = J_{Total} \int_0^{W_a} x \, dx
\]

\[
\Rightarrow J_n(W_a) - J_n(0) = J_T \int_0^{W_a} x \, dx
\]

At \( x = W_a \)

\( J_n(W_a) = J_T - J_{P_0} \)

At \( x = 0 \)

\( J_n(0) = J_{n_0} \)

Define a multiplication factor \( M \) as

\[
M = \frac{J_T}{(J_{n_0} + J_{P_0})} \quad \text{Total saturation current}
\]

\[
J_n(W_a) = J_T - J_{P_0}
\]

\[
\Rightarrow J_T - (J_{P_0} + J_{n_0}) = J_T \int_0^{W_a} x \, dx
\]

or

\[
1 - \frac{1}{M} = \int_0^{W_a} x \, dx
\]

\[
\Rightarrow M = \frac{1}{1 - \int_0^{W_a} x(E) \, dx}
\]

With the proper choice of field and \( W_a \) we can generate a diode breakdown condition.
\[ \int_0^\infty \alpha \, dx \to 1 \Rightarrow M \to \infty \]

Noting that \( M = \frac{1}{J_F} \), the above condition corresponds to a generated current which is \( \gg \frac{1}{J_F} \).

\[ J_F \to \infty \]
\[ \alpha M \to \infty \]

What does \( \int_0^\infty \alpha \, dx = 1 \) mean?

If \( E = \text{constant} \) in the avalanche region
\[ \alpha = \text{constant} \]
\[ \Rightarrow \alpha W_0 = 1 \]

Since \( \alpha = \# \text{ e-h pairs generated per unit distance per electron (or hole)} \), the above condition says that on the average each incoming electron (or hole) generates one electron-hole pair.
Consider an electron injected at $x = 0$. It drifts across the region $W_a$. If avalanche multiplication has to be sustained, the electron has to generate an e-h pair before it leaves the avalanche region. Assume that it generates an e-h pair just before it leaves the region. The hole drifts back across the avalanche region, and because $x = 0$ it generates an e-h pair near $x = 0$ before leaving the region. The generation process is self-sustaining when $M \to \infty$.

**Non Steady State Analysis**

**Continuity Equations**

\[
\frac{\partial n}{\partial t} = -\frac{1}{q} \frac{\partial H}{\partial x} + \alpha \frac{V}{s} (n+p)
\]

**Energy Equations**

\[
\frac{\partial P}{\partial t} = \frac{1}{q} \frac{\partial H}{\partial x} + \alpha \frac{V}{s} (n+p)
\]

Adding the equations together
Combining these two equations:

\[ I_s = \left(1 + \frac{I_C}{I_B}\right) I_B = I_B (1 + \beta) \quad (5.3) \]

From these relationships, it can be concluded that a large beta is desirable in order to operate the device with a high power gain. The device physical parameters that lead to a higher current gain are discussed in a later section of this chapter.

The common base configurations for the N-P-N and P-N-P bipolar transistors are shown in Fig. 5.2 together with the directions for the emitter, base, and collector currents. In this case, the base terminal serves as the reference. In the common base configuration, the input variable is the emitter current and the output variable is the collector current. Based upon Kirchhoff's law:

\[ I_s = I_C + I_B \quad (5.4) \]

The ratio of the collector current to the emitter current, called the common base current gain or \( \alpha \) (alpha) of the bipolar transistor, is given by:

\[ \alpha = \frac{I_C}{I_B} \quad (5.5) \]

The common emitter and common base current gains are related by:
Chapter 5: BIPOLAR TRANSISTORS

\[
\alpha = \frac{I_c}{I_g + I_c} = \frac{\beta I_g}{I_g + \beta I_g} = \frac{\beta}{1 + \beta} \tag{5.6}
\]

In a similar manner, it can be shown that:

\[
\beta = \frac{\alpha}{1 - \alpha} \tag{5.7}
\]

The device parameters that govern the current gain will be discussed in subsequent sections of this chapter.

5.2 CURRENT TRANSPORT

In both the common-base and common-emitter configurations, the collector current flow is controlled by the application of a voltage between the base and emitter terminals so as to forward bias the emitter-base junction. For the case of the NPN transistor, the forward bias results in the injection of electrons from the N⁺ emitter into the P-base region, and the injection of holes from the P-base region into the emitter region. The emitter current is comprised of both of these components. However, the portion of the emitter current due to hole injection into the N⁺ emitter region does not contribute to the collector current. The electrons injected into the P-base region diffuse through it and are collected at the reverse biased base-collector junction. Due to the finite minority carrier lifetime in the P-base region, some of the injected electrons are lost due to recombination.

The current transport between the emitter and collector for an NPN transistor in the common base configuration is shown in Fig. 5.3 with depletion boundaries and the internal current components indicated. The emitter-base junction is assumed to be under forward bias, while the collector-base junction is assumed to be under reverse bias. The current \(I_{eb}\) is the emitter current component due to electron injection at the emitter-base junction. The current \(I_{bc}\) is the electron current at the collector-base junction. Using these current components, the common base current gain can be written as:

\[
\alpha = \frac{\delta I_c}{\delta I_g} = \frac{\delta I_{eb}}{\delta I_g} \frac{\delta I_{bc}}{\delta I_{eb}} \tag{5.8}
\]

From this expression, it can be seen that the change in electron current flow between the emitter and collector can be partitioned into three portions. The first term, referred to as the \textit{emitter injection efficiency or gamma}, given by:

\[
\gamma_g = \frac{\delta I_{eg}}{\delta I_g} \tag{5.9}
\]
is the fraction of the total emitter current that comprises the electron current. As the name implies, it is a measure of the ability of the emitter to inject electrons into the base region in an efficient manner. Since the emitter current due to hole injection into the emitter region does not contribute to the collector current, the injection efficiency term is always less than unity.

The second term in Eq. (5.8), referred to as the base transport factor, given by:

\[ \alpha_T = \frac{\delta I_{nc}}{\delta I_{ne}} \]  \hspace{1cm} (5.10)

is a measure of the ability for electrons that are injected into the base from the emitter to reach the collector-base junction. In the presence of a finite recombination in the base region, some electrons will be lost during transport through the base region. Consequently, the base transport factor is always less than unity.

The third term in Eq. (5.8), referred to as the collector efficiency, given by:

\[ \gamma_c = \frac{\delta I_c}{\delta I_{nc}} \]  \hspace{1cm} (5.11)

is a measure of the ability for electrons to transport through the collector region. In the case of a reverse biased collector-base junction, a strong electric field is established within a depletion region at this junction. The electrons transported through the base region are swept out by this electric field into the collector region. At collector biases well below the avalanche breakdown voltage of the collector-base junction, this process occurs without loss of electrons. Consequently, the collector efficiency can be assumed to be equal to unity. However, at large collector biases, the electric field within the depletion region approaches the critical breakdown electric field. This results in the initiation of the avalanche multiplication process. In the
presence of avalanche multiplication, the collector current can exceed the electron current arriving at the collector-base junction. Under these conditions, the collector efficiency becomes greater than unity, and equal to the multiplication factor ($M$).

In the case of a power bipolar transistor, it is necessary to consider both the emitter injection efficiency and the base transport factor in analyzing the current gain. This is because, unlike signal transistors, the power transistors must be designed with relatively large base thickness to prevent reach-through breakdown. In addition, the operation of the power bipolar transistor at high current levels leads to high level injection in the base region, which has a strong impact on the current gain. These effects are described in the next section.

5.2.1 Emitter Injection Efficiency

In order to analyze the emitter injection efficiency, consider an NPN transistor with a narrow base region. As shown in Fig. 5.4, the emitter-base junction is forward biased while

![Diagram of an NPN transistor with labeled regions and carrier concentrations.](image)

**Fig. 5.4** Minority carrier distribution in a narrow base NPN transistor.

the collector-base junction is reverse biased. The applied forward bias across the emitter-base junction ($V_{be}$) produces injection of electrons into the base region and holes into the emitter region. Based upon the law of the junction, the injected concentration can be related to the equilibrium concentrations of the minority carriers ($n_{0B}$ and $p_{0B}$) on either side of the junction:
Chapter 5: BIPOLAR TRANSISTORS

\[ n_s(0) = n_{0s} \exp \left( \frac{q V_{ss}}{k T} \right) \]  
\[ p_s(0) = p_{0s} \exp \left( \frac{q V_{ss}}{k T} \right) \]

Due to the high doping concentration in the emitter, the diffusion length for the holes \( L_p \) is small. The continuity equation for holes in the emitter is given by:

\[ \frac{d^2 p}{dx^2} - \frac{p}{L_p^2} = 0 \]  

If the thickness of the \( N^+ \) emitter is much larger than the diffusion length, the hole concentration decays exponentially with distance from the junction to its equilibrium value as shown in Fig. 5.4. The solution for the hole distribution is given by:

\[ p(x) = p_s(0) \exp \left( -\frac{x}{L_p} \right) \]

with \( x \) increasing away from the junction edge. The hole current flowing at the emitter-base junction is then given by:

\[ J_p(0) = q D_{ps} \left[ \frac{dp}{dx} \right]_{x=0} \]  

Using Eqs. (5.15) and (5.16),

\[ J_p(0) = \frac{q D_{ps}}{L_{ps}} p_{0s} \exp \left( \frac{q V_{ss}}{k T} \right) \]

Now, consider the electron distribution in the base region. The continuity equation for electrons in the base is given by:

\[ \frac{d^2 n}{dx^2} - \frac{n}{L_e^2} = 0 \]

where \( L_e \) is the diffusion length for electrons in the base region. In the case of the narrow base transistor, the diffusion length is assumed to be much longer than the thickness \( W_e \) of the base region. Further, the reverse bias at the collector-base junction forces the electron concentration to go to zero at \( W_e \). Using this boundary condition, the electron concentration is found to decrease linearly from \( n_b(0) \) to 0 across the base region as shown in Fig. 5.4. Thus:
Chapter 5: BIPOLAR TRANSISTORS

\[ n(x) = n_b(0) \left(1 - \frac{x}{W_b}\right) \]  \hspace{1cm} (5.19)

The electron current flowing at the emitter-base junction is then given by:

\[ J_n(0) = q D_{nb} \left(\frac{dn}{dx}\right)_{x=0} \]  \hspace{1cm} (5.20)

Using Eqs. (5.19) and (5.20),

\[ J_n(0) = \frac{q D_{nb}}{W_b} n_{oE} \exp\left(\frac{q V_{BE}}{k T}\right) \]  \hspace{1cm} (5.21)

It is worth pointing out that these currents have been derived under the assumption that no recombination occurs in the base region. Consequently, all the base current supplies the recombination in the emitter region. The common emitter current gain, as determined by the emitter efficiency, is then given by:

\[ \beta_E = \frac{I_C}{I_E} = \frac{J_n(0)}{J_p(0)} \]  \hspace{1cm} (5.22)

Using Eqs. (5.17) and (5.21),

\[ \beta_E = \frac{D_{nb} P_{oE} L_{PE}}{D_{pe} P_{oE} N_b} \]  \hspace{1cm} (5.23)

It is important to note that, with the exception of the physical base width, all the other parameters are a strong function of the doping in the emitter and base regions. The dopant concentration dependence of the mobility (discussed in Chapter 2) must be taken into account when determining the diffusion coefficients. In the case of the minority carrier diffusion length in the emitter, in addition to the reduced diffusion coefficient, the effect of Auger recombination becomes very important in determination of the lifetime. Further, the equilibrium values of the minority carrier concentrations in the emitter and base regions must be determined after including the band gap narrowing phenomenon. If the intrinsic concentrations in the emitter and base regions are denoted as \( n_{iE} \) and \( n_{iB} \), respectively, the injection efficiency limited current gain can be related to the doping concentrations in the emitter \( (N_{oE}) \) and base \( (N_{oB}) \) regions:

\[ \beta_E = \left(\frac{D_{nb}}{D_{pe}}\right) \left(\frac{L_{PE}}{W_b}\right) \left(\frac{N_{oE}}{N_{oB}}\right) \left(\frac{n_{iE}^3}{n_{iE}^3}\right) \]  \hspace{1cm} (5.24)

It is desirable to design the power bipolar transistor to obtain a high current gain in order to reduce the base drive current. From the above equation, it can be concluded that this is
achievable by using a very low base doping concentration and a very high emitter doping concentration. However, this approach must be tempered by the following considerations:

(a) An increase in the emitter doping is accompanied by a reduction in the diffusion length due to Auger recombination and by an increase in the emitter intrinsic carrier concentration due to band gap narrowing. These effects counteract the increased doping concentration.

(b) A reduction in the base doping concentration leads to a low reach-through breakdown voltage and a poor output conductance due to depletion of the base region. Although these effects can be prevented by increasing the base width, this will also result in reduction of the current gain.

(c) A low base doping concentration results in high level injection in the base region at relatively low current densities. This produces a reduction in current gain as discussed in the next section.

(d) A low base doping concentration results in a high base sheet resistance which degrades current distribution under the emitter in the on-state and the storage time during turn-off.

The equation [Eq. (5.24)] for the emitter injection efficiency limited current gain was derived under the assumptions of a thick emitter region and no recombination in the base region. If the diffusion length for holes in the emitter is not much smaller than the emitter width ($W_e$), the holes injected into the emitter region can diffuse through it and reach the emitter (ohmic) contact. The effect of this upon the hole current can be taken into account by replacing the term $(D_{ps}, n_{eq}/I_{pe})$ by:

$$\int_0^{W_e} D_{ps}(x) \frac{n_{eq}^3(x)}{N_{eq}(x)} \exp\left(-\frac{x}{L_{ps}}\right) dx \quad (5.25)$$

An optimum doping level of about $1 \times 10^{19}$ per cm$^3$ for the emitter has been found to result in the highest gain for the bipolar transistor. The diffusion profile for the N$^+$ emitter region is usually empirically optimized to obtain the highest gain.

5.2.2 Emitter Efficiency Including High Level Injection in Base

The analysis in the previous section indicates that a low base doping level is desirable to obtain a high emitter injection efficiency. However, if the base doping level is low, the injected electron concentration in the base at higher current densities can become comparable to and even exceed the base doping level. This is referred to as high level injection in the base. Under high level injection conditions in the base, the majority carrier concentration increases in order to satisfy charge neutrality. The enhancement of the majority carrier concentration in the base produces an increase in the injection of holes into the N$^+$ emitter region. This results in a reduction in the injection efficiency and current gain of the transistor, which is referred to as the Rlitter effect.

Consider the emitter-base junction shown in Fig. 5.5 operating under high level injection conditions in the base region. High level injection is defined to take place when the injected minority carrier concentration exceeds the doping concentration. Since the doping concentration in the emitter region is very high, it will be assumed that low level injection conditions prevail in the emitter. In the base region, the majority carrier concentration is shown to exceed the
doping level ($p_{oB}$). Using the Boltzmann quasi-equilibrium boundary condition for the P-N junction:

$$\frac{p(x_p)}{n(x_p)} = \frac{n(x_p)}{n(x_p)} = \exp\left(-\frac{q\Delta\bar{\psi}}{kT}\right)$$  \hspace{1cm} (5.26)

where $\Delta\bar{\psi}$ is the potential barrier across the junction under the forward bias conditions. Note that under low level injection conditions:

$$n(x_p) = n_{oB}$$  \hspace{1cm} (5.27)

and

$$p(x_p) = p_{oB}$$  \hspace{1cm} (5.28)

Eq. (5.27) is still valid under high level injection conditions in the base because of the high doping concentration in the emitter, but Eq. (5.28) no longer holds true due to high level injection in the base region. Consequently,

$$p(x_p) \cdot n_{oB} = p(x_p) \cdot n(x_p)$$  \hspace{1cm} (5.29)

In order to satisfy the charge neutrality condition in the base region:

$$p(x_p) = p_{oB} - n_{oB} + n(x_p)$$  \hspace{1cm} (5.30)
Chapter 5: BIPOLAR TRANSISTORS

Since the equilibrium electron concentration in the base \((n_{0b})\) is small compared with the equilibrium majority carrier concentration in the base \((p_{0b})\):

\[
p(x_p) = p_{0b} + n(x_p)
\]  

(5.31)

Combining Eqs.(5.29) and (5.31) gives:

\[
p(x_p) = [p_{0b} + n(x_p)] \frac{n(x_p)}{n_{0b}}
\]  

(5.32)

According to the law of the junction:

\[
n(x_p) = n_{0b} \exp \left( \frac{q V_{BB}}{k T} \right)
\]  

(5.33)

where \(V_{BB}\) is the forward bias applied to the base-emitter junction. Combining this equation with Eq. (5.32) gives:

\[
p(x_p) = \frac{p_{0b}}{n_{0b}} \left[ 1 + \frac{n(x_p)}{p_{0b}} \right] n_{0b} \exp \left( \frac{q V_{BB}}{k T} \right)
\]  

(5.34)

The minority, majority, and intrinsic concentrations in the emitter and base regions are related by:

\[
p_{0b} \cdot n_{0b} = n_{1b}^2
\]  

(5.35)

and

\[
p_{0b} \cdot n_{0b} = n_{1b}^2
\]  

(5.36)

Using these relationships in Eq. (5.34):

\[
p(x_p) = p_{0b} \left( \frac{n_{1b}}{n_{1b}} \right)^2 \left[ 1 + \frac{n(x_p)}{p_{0b}} \right] \exp \left( \frac{q V_{BB}}{k T} \right)
\]  

(5.37)

In this expression, the term within the square brackets represents the increase in the injection of holes into the emitter region due to high level injection in the base region. If the electron injection into the base \([n(x_p)]\) is small compared with the base doping concentration \((p_{0b})\), this terms becomes equal to unity corresponding to the case for low level injection in the base region. The increase in hole injection into the emitter by the high level injection into the base degrades the injection efficiency.

An expression for the injection efficiency under high level injection conditions in the base can be derived by assuming that the injected carriers in the emitter decay exponentially with
Chapter 5: Bipolar Transistors

Distance away from the junction with a diffusion length $L_p$. Then:

$$P(x) = P(x_p) \exp \left( -\frac{x}{L_p} \right)$$ (5.38)

With $P(x_p)$ given by Eq. (5.37). The base current that supplies recombination current for the emitter is then given by:

$$I_B = \frac{A \alpha D_p x_p}{L_{ps}} P(x_p) \exp \left( \frac{q V_{bs}}{k T} \right)$$ (5.39)

If recombination in the base is neglected, the injected carrier concentration decreases linearly from $n(x_p)$ to zero across the base region as shown in Fig. 5.4. Consequently, the collector current is given by:

$$I_C = A \alpha D_{ns} \frac{n(x_p)}{W_B}$$ (5.40)

Using this equation:

$$n(x_p) = \frac{I_C W_B}{A \alpha D_{ns}} = \frac{J_C W_B}{q D_{ns}}$$ (5.41)

Combining these relationships, it can be shown that:

$$I_B = \frac{A \alpha D_p x_p}{L_{ps}} \left( \frac{n_{ib}}{n_{is}} \right)^2 P_{os} \left[ 1 + \frac{J_C W_B}{q D_{ns} P_{os}} \right] \exp \left( \frac{q V_{bs}}{k T} \right)$$ (5.42)

From Eq. (5.40), the collector current is given by:

$$I_C = \frac{A \alpha D_{ns} W_B}{P_{os}} n_{os} \exp \left( \frac{q V_{bs}}{k T} \right)$$ (5.43)

The common emitter current gain determined by the emitter injection efficiency is then obtained by taking the ratio of the collector to the base current:

$$\beta = \frac{L_{ps} D_{ns} n_{os}}{W_B D_{ps} P_{os} \left( 1 + \frac{J_C W_B}{q D_{ns} P_{os}} \right)}$$ (5.44)

This equation indicates that, at high current densities, the current gain will decrease inversely with increasing current density. A schematic illustration of this behavior of the current gain is provided in Fig. 5.6, where the current gain and current density have been plotted on a logarithmic scale. A fall-off in the current gain with increasing collector current is highly
Fig. 5.6 Variation of current gain due to high level injection in the base.

disadvantageous because the base drive current for the transistor increases as the square of the collector current.

The current density at which the current gain decreases by a factor of 2 due to high level injection in the base has been referred to as the Rittner current density ($J_R$). Based upon this definition:

$$\beta = \frac{\beta_0}{[1 + (J_C / J_R)]} \quad (5.45)$$

An expression for the Rittner current density can be obtained from Eq. (5.44):

$$J_R = \frac{q D_n^* P_{0B}}{N_p} \quad (5.46)$$

It is advantageous to obtain a high value for the Rittner current density because this moves the fall-off in current gain to higher current densities. It can be seen that the Rittner current density can be increased by increasing the base doping concentration ($p_n$) and decreasing the base width ($W_B$). However, an increase in the base doping concentration decreases the current gain at low current levels, resulting in no improvement in the overall gain. Any decrease in the base width must be done after taking into account its impact on the reach-through breakdown voltage.
5.2.3 Base Transport Factor

The base transport factor is a measure of the ability of the injected carriers to get from the emitter/base junction to the base/collector junction. Since the transport of the carriers occurs by diffusion through the base region, the base transport factor ($\alpha_T$) is given by:

$$\alpha_T = \frac{x_{nc}}{x_{m}} = \frac{[\delta n / \delta x]_{x \rightarrow x_{n}}}{[\delta n / \delta x]_{x \rightarrow 0}}$$  \hspace{1cm} (5.47)

If recombination in the base region is negligible:

$$\left[ \frac{\delta n}{\delta x} \right]_{x \rightarrow x_{n}} = \left[ \frac{\delta n}{\delta x} \right]_{x \rightarrow 0}$$  \hspace{1cm} (5.48)

and the base transport factor becomes equal to unity. In this case the minority carrier concentration decreases linearly from the emitter to the collector junction.

When the recombination in the base region cannot be neglected, the carrier concentration

![Fig. 5.7 Distribution of injected carriers in the base region.](image)

in the base region no longer decreases linearly as shown in Fig. 5.7. In order to analyze the carrier distribution, consider the continuity equation for electrons in the steady state:

$$\frac{d^2 n}{dx^2} - \frac{n}{L_a^2} = 0$$  \hspace{1cm} (5.49)

The solution for this equation has the form:
Chapter 5: BIPOLAR TRANSISTORS

\[ n = A e^{-x/L_B} + B e^{x/L_B} \]  \hspace{1cm} (5.50)

where the constants A and B are determined by the boundary conditions:

\[ n = n_B(0) = n_{0B} e^{qV_{BB}/kT} \]  \hspace{1cm} (5.51)

at \( x = 0 \), and

\[ n = 0 \]  \hspace{1cm} (5.52)

at \( x = W_B \). The solution for the carrier distribution in the base is then obtained as:

\[ n = n_{0B} \frac{\sinh[(W_B - x)/L_B]}{\sinh(W_B/L_B)} \exp\left(\frac{qV_{BB}}{kT}\right) \]  \hspace{1cm} (5.53)

By using this carrier distribution to obtain \( \delta n/\delta x \) at the emitter and collector ends of the base region, the base transport factor can be obtained:

\[ \alpha_T = \frac{1}{\cosh(W_B/L_B)} \]  \hspace{1cm} (5.54)

If the diffusion length \( (L_B) \) is much larger than the base width \( (W_B) \), the base transport factor can be calculated using the following approximation:

\[ \alpha_T = 1 - \frac{W_B^2}{2L_B^2} \]  \hspace{1cm} (5.55)

This expression indicates that a higher current can be obtained by reducing the base width and maintaining a large minority carrier diffusion length in the base.

5.2.4 Collector Bias Effects

The above analysis of the base transport factor was performed under the assumption that the base/collector junction is reverse biased. The current gain determined by the base transport factor is then dependent upon the undepleted base width \( (W_B) \). When the collector bias is increased, the voltage is supported across a larger depletion region thickness. In the case of a bipolar power transistor, a lightly doped drift region is used in the collector region to support high collector voltages. In spite of the lightly doped collector drift region, a portion of the collector bias is also supported across a depletion region in the base region. The thickness of the depletion region in the base increases with increasing collector bias. This results in a reduction in the thickness of the undepleted base width. As a consequence, the base transport
factor increases with increasing collector bias. This produces a decrease in the output conductance of the bipolar power transistor.

A typical shape for the output characteristics of a bipolar transistor at large collector biases is illustrated in Fig. 5.8. As an approximation, the collector current is shown to increase linearly with increasing collector voltage. An extrapolation of all the output characteristics to zero collector current is indicated to lead to a common intercept on the x-axis. The voltage at which this intercept occurs is referred to as the Early Voltage ($V_\text{EB}$). In actual devices, the output characteristics are not linear particularly at higher collector biases. This behavior can be analyzed by considering depletion of the base region at higher collector voltages.

Consider the N-P-N bipolar power transistor illustrated in Fig. 5.9 at two collector bias voltages $V_1$ and $V_2$, with the bias $V_1$ larger than the bias $V_2$. The thickness of the depletion region in the base is related to the base doping concentration $N_{AB}$ and the voltage supported in the base region ($V_B$) by:

$$W_B = \sqrt{\frac{2 e V_B}{q N_{AB}}} \quad (5.56)$$

Due to the relatively low doping concentration in the N-drift region, most of the collector bias is supported across the drift region. The voltage supported across the base region can be determined by taking into account the difference in the doping concentrations in the base and drift regions:
Chapter 5 : BIPOLAR TRANSISTORS

Fig. 5.9 Effect of increasing collector bias on carrier profile in the base region.

\[ V_B = V_C \left( \frac{N_{Dr}}{N_{Ae} + N_{Dr}} \right) \]  \hspace{1cm} (5.57)

where \( N_{Dr} \) is the doping concentration in the N-drift region. It should be noted that either a lower doping concentration in the collector or a higher doping concentration in the base region will result in less voltage being supported across the base region. The undepleted base thickness is then given by:

\[ W = W_0 - W_D = W_0 - \frac{2 \mu_{n} V_B}{q N_{Ae}} \]  \hspace{1cm} (5.58)

Assuming that recombination in the base region is negligible, the collector current is given by:

\[ J_c = q D_n \frac{n_{oB}(0)}{W} \]  \hspace{1cm} (5.59)

Combining Eq. (5.58) and (5.59):
Chapter 5: BIPOLAR TRANSISTORS

\[ J_c(V_{CB}) = qD_n \frac{n_s(0)}{N_s} \frac{1}{[1 - \sqrt{2 e V_B} / (q N_{AB} N_s^2)]} \] (5.60)

At low collector voltages, the depletion width in the base region is small. Under these conditions, the collector current is given by:

\[ J_{co} = qD_n \frac{n_s(0)}{N_s} \] (5.61)

For a good transistor design, the quantity inside the square root is usually small. By using a series expansion for the square root and neglecting higher order terms, it can then be shown that:

\[ J_c = J_{co} \left[ 1 + \frac{e V_B}{q N_{AB} N_s^2} \right] \] (5.62)

Using Eq. (5.57):

\[ J_c = J_{co} \left[ 1 + \frac{e}{q N_{AB} N_s^2} \frac{V_{CB}}{1 + (N_{AB}/N_{c})} \right] \] (5.63)

This equation indicates that the collector current will increase linearly with increasing collector voltage. Using this equation, the Early voltage, at which the collector current extrapolates to zero, can be determined:

\[ V_{e} = \frac{q N_{AB} N_s^2}{e} \left[ 1 + \frac{N_{AB}}{N_{c}} \right] \] (5.64)

The Early voltage predicted by this equation is independent of the collector current at low biases, which implies that all the output characteristics should extrapolate to a common intercept as illustrated in Fig. 5.8. A large value for the Early voltage is desirable to obtain good output characteristics. From Eq. (5.64), it can be concluded that this is achievable by using a base region with large width and doping concentration. Unfortunately, this is in conflict with obtaining a high current gain.

5.2.5 Voltage Saturation Region

In the previous section, the collector bias was assumed to be sufficiently large to maintain the collector junction under reverse bias. This does not hold true when the collector voltage is reduced to values approaching the base drive voltage (about 1 volt). When the collector base
junction becomes forward biased, the minority carrier profile in the base region becomes altered from that shown in Fig. 5.4. Since both the emitter-base junction and the collector-base junction are forward biased, the minority carrier profile can be obtained by using the superposition principle. This assumes reciprocity between the emitter-base and collector-base junctions. The superposition principle is valid only if the system is linear. Although the bipolar transistor is highly non-linear in terms of relating the terminal currents and voltages, it is fortunately linear in terms of relating the injected minority carrier concentration in the base to the terminal currents. Based upon these considerations, the minority carrier profile in the base region can be obtained by using the minority carrier profiles for operation of the transistor in the forward active region and the reverse active region. The forward active region is the case of a forward biased emitter-base junction and a reverse biased collector-base junction as discussed in the previous section. The reverse active region is the case of a forward biased collector-base junction and a reverse biased emitter-base junction. The minority carrier profiles for these cases are compared in Fig. 5.10. Superposition of these carrier profiles provides the carrier profile in the voltage saturation region. It can be seen that the slope of the carrier profile is reduced when the collector-base junction becomes forward biased. The collector current is then also reduced:

\[
J_c = \frac{n_b(0) - n_b(W_p)}{W_p} \tag{5.65}
\]
Chapter 5: BIPOLAR TRANSISTORS

From the above discussion, the on-state voltage drop across the bipolar transistor in the voltage saturation region is less than the junction potential (typically 0.7 volts). This is an advantage from the point of view of reducing the power dissipation in the on-state. It is important to note that this voltage drop is the internal voltage drop from the emitter-base junction to the base-collector junction. It does not include the voltage drop across the N-drift region in the power bipolar transistor. The doping profile of a power bipolar transistor is illustrated in Fig. 5.11 for the case of a device fabricated using emitter and base diffusion into

![Doping profile for an epitaxial power bipolar transistor.](image)

Fig. 5.11 Doping profile for an epitaxial power bipolar transistor.

a high resistivity epitaxial layer grown on a highly doped substrate. The thickness of the epitaxial layer is chosen to support large collector voltages.

The specific resistance of the collector drift region can be related to its resistivity ($\rho_D$) and thickness ($W_D$):

$$R_{D,e} = \rho_D W_D = \frac{N_D}{Q \mu_n N_D}$$

(5.66)

Due to the relatively high resistivity of the N-drift region (typically 50 ohm-cm) and its large thickness (typically 50 microns), the specific resistance of the drift region is large (typically 0.25 ohm-cm$^2$). This would result in a high voltage drop across the drift region at typical on-state current density in the collector (50 amperes per cm$^2$). Fortunately, in the case of the power bipolar transistor, this resistance is reduced by the injection of minority carriers from the base region into the drift region. Due to the low doping concentration in the drift region, high level injection occurs, resulting in conductivity modulation of the drift region. The enhancement in the conductivity of the drift region by the injected carriers results in a reduction in the on-state voltage drop. When the collector bias is sufficiently small, the base-collector junction operates
with a sufficient forward bias to produce high level injection of carriers throughout the drift region. This regime of operation is called the voltage saturation region. As the collector bias is increased at any fixed base drive current, the forward bias across the base-collector junction is reduced. This results in a decrease in the injection of carriers into the drift region. Under these conditions, only a portion of the drift region near the base-collector junction operates under high level injection while the rest of the drift region is not modulated. This regime of operation is called the voltage quasi-saturation region.

![Diagram](image)

**Fig. 5.12** Output characteristics of a power bipolar transistor showing the saturation and quasi-saturation regions of operation.

The different operating regions are illustrated in Fig. 5.12 at relatively small collector voltages close to the on-state voltage drop. Note that the current gain of the transistor is decreasing when proceeding from the forward active region through the quasi-saturation region into the saturation region. This is because of the additional recombination current that must be supplied by the base electrode for sustaining carriers that are injected into the collector drift region. Consider the bias points shown in Fig. 5.12 for a fixed base drive current. From 0 to A, the device is in the voltage saturation region with the conductivity of the entire drift region modulated by the injected carriers. From A to B, the device is in the quasi-saturation region. At collector voltages above bias point A, the conductivity of a part of the drift region is no longer modulated by the injected carriers, resulting in an additional voltage drop across the device. At collector voltages above bias point B, the device enters the forward active region. The bias point B is the point of transition from forward biased to reverse biased operation for the base-collector junction.

The analysis of the saturation and quasi-saturation regions can be performed by determination of the minority carrier distribution profile in the drift region. Consider the cross-section of the power bipolar transistor shown in Fig. 5.13 with the minority carrier distribution...
Fig. 5.13 Minority carrier distribution in the collector drift region under saturation and quasi-saturation operation.

in the drift region indicated for the case of the saturation and quasi-saturation regions. The base-collector junction is assumed to be forward biased resulting in the injection of minority carriers (holes) into the drift region. In this figure, the minority carrier concentration at the base-collector junction is defined as \( p(0) \) for both cases. In the drift region, charge neutrality demands:

\[
\frac{dn}{dx} = \frac{dp}{dx} \tag{5.68}
\]

Due to the low doping concentration in the drift region to obtain high blocking voltage capability, high level injection conditions can be assumed leading to:

\[
J_{pc} = q \mu_p p E - q D_p \frac{dp}{dx} \tag{5.69}
\]
Chapter 5: BIPOLAR TRANSISTORS

\[ J_{sc} = q \mu_n n E + q D_n \frac{dn}{dx} \]  \hspace{1cm} (5.70)

Under saturation and quasi-saturation conditions, the base current supplies recombination in the emitter, base and collector regions as indicated in Fig. 5.13. The recombination currents supplied to the emitter and base regions have been previously discussed. The recombination current supplied to the collector drift region is represented by \( J_{sc} \). If the current gain is assumed to be large, then this current can be assumed to be small (i.e., approximately equal to zero). Then, from Eq. (5.69):

\[ E = \frac{kT}{q} \frac{1}{p} \frac{dp}{dx} \]  \hspace{1cm} (5.71)

Further, since the hole current is assumed to be small, the electron current (\( J_c \)) can be assumed to be equal to the total collector current (\( I_c \)). By using Eq. (5.71) in Eq. (5.70) with the relationships in Eqs. (5.67) and (5.68), it can be shown that:

\[ J_c = q \mu_n (p + N_p) \left( \frac{kT}{q} \frac{1}{p} \frac{dp}{dx} \right) + q D_n \frac{dp}{dx} \]  \hspace{1cm} (5.72)

Making use of the Einstein relationship between the mobility and diffusion coefficient:

\[ J_c = 2qD_n \left( 1 + \frac{N_p}{2p} \right) \frac{dp}{dx} \]  \hspace{1cm} (5.73)

The solution of this equation, with the boundary condition that the hole concentration at the base-collector junction is \( p(0) \), gives the minority carrier distribution in the drift region:

\[ p(x) = p(0) - \frac{J_c x}{2qD_n} + \frac{N_p}{2} \ln \left( \frac{p(0)}{p(x)} \right) \]  \hspace{1cm} (5.74)

Due to the low doping concentration in the drift region, the last term in Eq. (5.74) can be neglected. Thus:

\[ p(x) = p(0) - \frac{J_c x}{2qD_n} \]  \hspace{1cm} (5.75)

This equation indicates a linear decrease in the minority carrier concentration from \( p(0) \) at the base-collector junction as illustrated in Fig. 5.13 for both the saturation and quasi-saturation cases. In the saturation case, the injected minority carrier concentration remains much larger than the majority carrier concentration (\( N_p \)) throughout the drift region. In the quasi-saturation case, the minority carrier concentration remains much higher than the majority carrier concentration over only a portion (\( W_m \)) of the drift region. This distance (\( W_m \)) can be determined by assuming that the minority carrier concentration becomes equal to the equilibrium
Chapter 5: BIPOLAR TRANSISTORS

value at this position:

$$P(N'_N) = \frac{n_i^2}{N_D}$$  \hspace{1cm} (5.76)

Substituting this into Eq. (5.75) gives:

$$W'_N = \frac{2 q D_a (P(0) - N_D)}{J_c} = \frac{2 q D_a P(0)}{J_c}$$  \hspace{1cm} (5.77)

The voltage drop in the collector drift region can be obtained by integration of the electric field distribution as determined by Eq. (5.71):

$$V_M = -\int_{x_0}^{x_M} E \, dx = -\frac{k T}{q} \int_{P(0)}^{P(x_M)} \frac{dp(x)}{p(x)}$$  \hspace{1cm} (5.78)

Using Eq. (5.75):

$$V_M = \frac{k T}{q} \ln \left( \frac{P(0)}{N_D} \right)$$  \hspace{1cm} (5.79)

The magnitude of this voltage drop in the modulated region is relatively small (typically between 100 and 200 millivolts). Since the modulated region extends throughout the drift region under saturation conditions, the on-state voltage drop across the power bipolar transistor is also small under voltage saturation conditions.

At the onset of quasi-saturation (point A in Fig. 5.12), the thickness of the modulated region becomes equal to the total drift region thickness ($W_D$). The collector current density at which quasi-saturation begins to occur can then be obtained by using Eq. (5.77) with $W'_N$ equal to $W_D$:

$$J_{Ces} = \frac{2 q D_a P(0)}{W_D}$$  \hspace{1cm} (5.80)

In the quasi-saturation region, an ohmic voltage drop occurring in the unmodulated portion of the drift region must be added to the voltage drop discussed above for the modulated portion. The voltage drop in the unmodulated portion of the drift region is given by:

$$V_U = \frac{J_c (W_D - W'_N)}{q \mu_n N_D}$$  \hspace{1cm} (5.81)

The stored charge per unit area in the collector drift region due to high level injection is obtained by integration of the minority carrier profile:
Chapter 5: BIPOLAR TRANSISTORS

\[ Q_{sp} = \frac{q \mu(0) W_N}{2} \]  

(5.82)

and the base current supplied to support recombination in the collector drift region is given by:

\[ J_{pc} = \frac{q \mu(0) W_N}{2 \tau_{HL}} \]  

(5.83)

This current is responsible for the reduced current gain observed in the saturation and quasi-saturation regions.

The saturation and quasi-saturation regions are important because the bipolar power transistor is operated in these regions during the on-state. From the point of view of reducing the power dissipation in the on-state, it is desirable to operate the transistor in the saturation region. However, as shown above, the stored charge in the collector drift region becomes much larger when the transistor is driven into the saturation region. During turn-off, this stored charge must be removed before the transistor can begin to support voltage. A larger stored charge in the collector drift region results in a longer storage phase during turn-off for a fixed reverse base drive. This is undesirable when the transistor must be operated at higher frequencies.

5.2.6 Base Widening at High Current Densities

Another phenomenon of importance to power bipolar transistors is an increase in the effective base width through which the carriers must diffuse when the device is operated at high collector current densities. This phenomenon occurs in the forward active region of operation where the base-collector junction is reverse biased. The high collector current density in the drift region is supported by the drift of the majority carriers under the influence of a large electric field created by the large collector bias. Due to the high electric fields prevalent in the drift region, the majority carriers can be assumed to be moving at their saturated drift velocity \( v_s \), which is approximately \( 10^7 \) cm/sec. When the collector current density is large, there is a correspondingly large majority carrier density in the drift region given by:

\[ n = \frac{J_c}{q v_s} \]  

(5.84)

The presence of these majority carriers (electrons) alters the electric field distribution because their charge (negative) subtracts from the charge (positive) due to the donors. The Poisson’s equation for the collector depletion region is then:

\[ \frac{d\phi(x)}{dx} = -\frac{\mathcal{L}}{\varepsilon_s} [N_D - n] \]  

(5.85)
Substituting for $n$ from Eq. (5.84) and performing an integration:

$$ E(x) = E(0) - \frac{q}{\varepsilon_s} \left[ N_D - \frac{J_C}{q V_s} \right] x $$

(5.86)

It can be seen that the electric field in the drift region varies linearly with distance as in the case of a reverse biased junction but its rate of variation is dependent upon the current density. At low current densities, the electric field distribution is the same as that for the case of static blocking conditions with the peak of the electric field located at the base-collector junction (i.e.

![Diagram of bipolar transistor with electric field distribution](image)

**Fig. 5.14** Electric field distribution in the collector drift region at high current densities.

At $x = 0$. This is indicated by the curve labeled 'a' in Fig. 5.14. As the current density increases, the contribution from the majority carriers in the drift region becomes significant and, for a fixed applied collector bias, the slope of the electric field profile becomes smaller as shown by curve 'b'. It should be noted that, for the case of the epitaxial power bipolar transistor with a highly doped substrate, the depletion region is truncated by the high doping concentration in the N+ substrate. As the current density increases, a value is reached at which the slope of the electric field profile becomes zero as shown by curve 'c'. This current density is given by:

$$ J_{co} = \frac{q V_s N_D}{2} $$

(5.87)
Chapter 5: Bipolar Transistors

At this current density, the number of electrons per cm$^2$ moving through the collector drift region becomes exactly equal to the doping concentration in the drift region. At even higher current densities, the electron density in the drift region exceeds the doping concentration. This produces a change in the polarity of the net charge from positive to negative, which reverses the slope of the electric field profile as shown in curve 'd'. The peak of the electric field no longer occurs at the base-collector junction but now occurs at the interface between the lightly doped drift region and the highly doped substrate. At a sufficiently large current density, the electric field at the base-collector junction can become equal to zero. This condition is shown in Fig. 5.14 as curve 'e'. The current density at which this takes place is called the Kirk current density. Solution of Poisson's equation for this case gives:

$$J_x = q \nu_s N_D + \frac{2 \epsilon_s V_{ce}}{W_N}$$  \hspace{1cm} (5.88)

This current density is important because, at even higher values, the electric field profile shifts

![Diagram of electric field distribution](image)

**Fig. 5.15** Electric field distribution at high current densities illustrating the formation of a current induced base region within the drift region.

to that shown in Fig. 5.15 as curve 'f'. Note that, in this case, there is an undepleted portion of the drift region on the right hand side of the base-collector junction. This undepleted portion of the drift region is referred to as the current induced base region because it is formed by the
Chapter 5: BIPOLAR TRANSISTORS

presence of a high collector current density flowing through the drift region and because the injected electrons from the emitter-base junction must diffuse not only through the physical base width \(W_b\) but must traverse the undepleted portion of the drift layer \(W_{cb}\) before they can be swept out by the high electric field within the rest of the drift region of width \(W_d\). This increase in the effective base width leads to a reduction in the current gain at higher current densities.

As indicated by Eq. (5.88), the formation of the current induced base is related to the device drift region parameters. For the case of bipolar power transistors with lower breakdown voltages, the drift region doping \(N_D\) is relatively large and its width \(W_{cb}\) is small. This leads to the first term in Eq. (5.88) becoming dominant and the Kirk current density becomes independent of the collector bias \(V_{cb}\). On the other hand, for the case of bipolar transistors with high breakdown voltages, the drift region doping \(N_D\) is relatively small and its width \(W_{cb}\) is large. This leads to the second term in Eq. (5.88) becoming dominant and the Kirk current density becomes strongly dependent on the collector bias \(V_{cb}\). This is illustrated in Fig. 5.16

![Graph showing Kirk current density dependence on collector-base voltage](image)

**Fig. 5.16** Dependence of the Kirk current density on collector-base bias.

for three cases of drift region doping.

The width \(W_{cm}\) of the current induced base region can be determined by using the electric field distribution illustrated in Fig. 5.15. This electric field distribution occurs at high current densities. Consequently, the majority carrier density in the drift region given by Eq. (5.84) is much larger than the doping concentration in the drift region. Under these conditions, the electric field distribution given by Eq. (5.86) can be rewritten as:
\[ E(x) = \left( \frac{J_c}{e_s v_s} \right) x \]  

(5.89)

Since the doping concentration in the N+ substrate is much larger than the majority carrier concentration in the drift region, the collector-base voltage can be assumed to be supported across a width \( W_b \), located primarily within the drift region. From Eq. (5.89), this width is related to the applied bias by:

\[ W_b = \sqrt{\frac{2 e_s v_s V_{cb}}{J_c}} \]  

(5.90)

Thus, the current induced base width is given by:

\[ W_{CB} = W_b - \sqrt{\frac{2 e_s v_s V_{cb}}{J_c}} \]  

(5.91)

and the effective base width of the transistor at high current densities is given by:

\[ W_b(J_c) = W_b + W_b - \sqrt{\frac{2 e_s v_s V_{cb}}{J_c}} \]  

(5.92)

The current gain of the bipolar transistor is reduced because the electrons injected from the emitter must traverse through this increase in the effective base width. An estimate of this phenomenon can be obtained by using Eq. (5.55) to derive an expression for the common emitter current gain:

\[ \beta = \frac{2 L_{ns}^2}{W_b^2} \]  

(5.93)

By substituting for \( W_b \) from Eq. (5.91), it can be shown that:

\[ \beta = \beta_{J=0} \left[ 1 - 2 \left( \frac{W_b - \sqrt{2 e_s v_s V_{cb}/J_c}}{W_b} \right) \right] \]  

(5.94)

under the assumption that the base widening is small compared with the physical base width \( (W_b) \). Here, \( \beta_{J=0} \) is the current gain at low current levels. This expression indicates that the current gain decreases with increasing current density, while it increases with increasing collector-base bias voltage. A physical insight into this behavior can be obtained by considering the electric field profiles shown in Fig. 5.17. When the current density increases, the majority carrier density in the drift region increases leading to change in the slope of the electric field
Chapter 5: BIPOLAR TRANSISTORS

Fig. 5.17 Electric field profiles illustrating effect of (a) current density and (b) collector-base voltage up on the current induced base formation.

profile. This produces an increase in the width of the current induced base region. In contrast to this, when the collector-base voltage is increased, the area under the electric field profile increases, leading to a reduction in the width of the current induced base region.

It was previously demonstrated that the current gain of the bipolar transistor decreases when the current density is increased due to onset of high level injection in the base region (Ritter effect). The formation of the current induced base region (Kirk effect) occurs at even

Fig. 5.18 Reduction in current gain of a bipolar transistor with increasing collector current density.
higher current densities leading to an enhancement in the degradation of current gain with increasing current density. This is illustrated in Fig. 5.18.

3.2.7 Emitter Current Crowding

In the previous sections, the current flow through the bipolar power transistor was assumed to occur uniformly through its entire emitter area. This forms the basis for the one-dimensional analysis used to derive all the expressions. In an actual power transistor, the emitter and base contacts must be interdigitated as illustrated in Fig. 5.19. When the transistor is biased into its on-state, a base drive current is applied at the base terminal. This current flows across the forward biased emitter-base junction resulting in the injection of electrons from the emitter. Due to the interdigitation of the contacts, in order to forward bias the emitter-base junction at point A, the base current must flow through the base resistance $R_b$. This produces a voltage drop between points A and B resulting in the forward bias across the emitter-base junction at point A being lower than that at point B. Consequently, the emitter injection at point A is smaller than that at point B and the emitter current distribution is not uniform across the width of the emitter N+ diffusion.

The resistance of the base region through which the base current flows is determined by the pinch-resistance formed by the superposition of the base and emitter diffusions. The integrated net P-type doping concentration in the base region below the emitter is much smaller than the P-type dopant diffused into the device. This can be seen from the doping profiles under the emitter shown in Fig. 5.20. Further, the depletion layers formed in the base region, at both
GENERAL THEORY FOR PINCHED OPERATION OF THE JUNCTION-GATE FET*

A. B. GREBENE
Signetics Corporation, Research and Development Lab., Sunnyvale, California, U.S.A.

and

S. K. GHANDHI
Division of Electrophysics, Rensselaer Polytechnic Institute, Troy, New York, U.S.A.

(Received 21 June 1968; in revised form 16 December 1968)

Abstract—A device oriented model is developed to describe the operation of the junction-gate field-effect transistor (FET) beyond pinch-off. The model is derived on the basis of a generalized structure with an arbitrary channel doping profile. It provides a qualitative and quantitative description of the current conduction mechanism, and is applicable over the entire dynamic range of device operation. Current conduction mechanisms in the vicinity of the source and the drain are examined separately. It is shown that the saturation of carrier drift velocities at high electric fields results in formation of a drain space-charge region of finite length. An approximate solution of the two-dimensional Poisson’s equation is developed to describe the potential distribution within this region. A significant result of the device model is the prediction of a finite drain resistance in pinched operation, which shows a strong dependence on the device operation point.

Résumé—Un modèle orienté à un dispositif est développé pour décrire l’opération du transistor à effet de champ (FET) à partir de la zone de la pointe. Le modèle est dérivé sur la base d’une structure généralisée ayant un profil de la source de canal arbitraire. Il fournit une description qualitative et quantitative du mécanisme de conduction et est applicable à l’entière gamme d’opérations du dispositif.

Les mécanismes de conduction de courant aux environs de la source et du drain sont examinés séparément. On démontre que la saturation des vitesses d’aspiration de porteurs aux champs électriques élevés résulte en la formation d’une région de charge d’espace de drain de longueur limitée. Une solution approximative de l’équation de Poisson à deux dimensions est développée pour décrire la distribution de potentiel dans la région. Un des résultats les plus significatifs du modèle est la prédiction d’une résistance de drain limitée dans l’opération pinçée, démontrant une forte dépendance sur le point opérationnel du dispositif.


*This work was supported by the U.S. Army Research Office, Durham, N.C. on ARO-D Contract No. DAHC04 67C 0035.

573
INTRODUCTION

The behavior of the junction-gate field-effect transistor (FET) at values of the drain potential less than the pinch-off voltage can be accurately described by means of a first-order "gradual-channel" approximation. This model was originally developed by Shockley\(^{11}\) for a uniformly-doped channel structure, and later expanded to devices with arbitrary impurity distribution.\(^{2}\)

The gradual-channel model is based on the assumption that the electric field gradients within the depleted channel region are normal to the direction of current flow. For operation with drain-to-gate voltages in excess of the device pinch-off voltage ("pinched-mode" operation), this model is no longer valid, particularly in the vicinity of the drain terminal. It is commonly accepted practice to describe device behavior beyond pinch-off by an arbitrary extrapolation of its V–I characteristics outside of their region of validity. This assumes complete independence of the output characteristics on the source-drain voltage, as well as zero source-drain conductance in the pinched mode. Therefore, the model fails to offer a physically acceptable description of the current conduction mechanism through the FET channel for device operation beyond pinch-off.

The exact determination of the FET characteristics for pinched operation requires the solution of the two-dimensional Poisson's equation under extremely complex boundary conditions. Grovaxelet and Jund\(^{3}\) have shown that, under certain idealized assumptions, it is possible to obtain a computer solution to this equation over the entire channel region by an iterative numerical approximation technique. However, their solution fails to give a physical insight to phenomena responsible for device operation. In addition, it is not readily applicable to practical problems of device design and characterization since its results and predictions cannot be expressed in a closed form.

In his original paper,\(^{11}\) Shockley suggested the possible use of a two-region model to describe the current conduction mechanism in pinched operation. This approach permits the separation of the FET channel into two regions in the direction of current flow, and allows the independent examination of conduction phenomena in the vicinity of the source and the drain terminals. An approximate solution for the potential and the electric field distribution over the entire channel is next obtained by matching the solutions in each of these regions at a predetermined interface within the channel, normal to the direction of current flow.

Various mathematical methods of joining these solutions have been discussed by Shockley,\(^{11}\) by Prim and Shockley,\(^{14}\) and more recently by Wu and Sah.\(^{15}\) These models share two common features:

(a) Matching of the solutions in the two regions of the device is accomplished by a curve-fitting procedure, in terms of two adjustable parameters whose values are chosen to obtain the best match.

(b) The resulting device model is a mathematical one, and does not offer a physical description of the current conduction mechanism through the pinched channel region near the drain.

The aim of this paper is to develop a device-oriented physical model for the junction-gate FET which avoids the shortcomings of these mathematical models. The resulting model provides a working description of the V–I characteristics and a.c. parameters in pinched operation, in terms of the device geometry, the doping profile, and the electrical properties of the semiconductor material.

The characteristics of the FET beyond pinch-off are analyzed by treating the device channel as made up of two separate regions along the direction of current flow. In Region I, in the vicinity of the source, the gradual-channel approximation is used, and the depletion layer profile estimated from a solution of the one-dimensional Poisson's equation. In Region II, near the drain, the two-dimensional Poisson's equation is solved to determine the potential distribution. The boundary values for the solution of Poisson's equation within Region I and II are chosen to require the continuity of the potential distribution across the interface separating them.

It is shown that the current flow is constrained in Region II to a narrow conductive filament (residual channel) along the longitudinal axis of symmetry of the metallurgical channel. For practical device geometries, the saturation of carrier drift velocity at high electric field strengths is found to be the dominant factor in determining the dimensions of this region. Consequently, in
the development of the model, it is assumed that the interface between the two regions of the channel is determined by the onset of carrier drift velocity saturation.\textsuperscript{(16 - 18)} The validity of this assumption, as well as other possible conduction mechanisms that may contribute to the total drain current in pinched-mode operation, is examined closely prior to the development of the model.

In the following analysis, an \textit{n}-channel FET structure with parallel gate regions and an arbitrary channel impurity profile is considered. It is assumed that the gate regions are heavily doped, so that the gate-channel depletion layer spreads predominantly into the channel. Each of the gate regions is assumed to be independently biased with respect to the source. The structural diagram and the channel depletion layer profile for such an FET in pinched operation are shown schematically in Fig. 1. Since most commercially available units voltage, \( V_p \), the drain end of the channel is totally depleted of mobile carriers. Since the drain current, \( I_d \), is finite, this results in a physically unrealizable situation, requiring current conduction through a region of the channel that is totally devoid of mobile carriers.

A number of mechanisms may be postulated to explain current flow through a totally depleted channel section. These are as follows:

(a) Avalanche multiplication of carriers within the depleted channel section (Region II). The onset of this mechanism should result in a sharp increase of the noise level associated with the drain current in pinched-mode operation; this is not observed in practice. Furthermore, electric field strengths encountered within the drain space-charge layer (see Fig. 4) are typically well below those values at which avalanche multiplication effects become significant.

(b) Gate-channel junction breakdown, resulting in lateral carrier flow from gate to the drain. This can only occur when the reverse breakdown voltage of the gate-drain junction is exceeded. This voltage is typically far in excess of the pinchoff voltage, \( V_{p} \). In addition, breakdown of this type would result in a sudden decrease of gate impedance accompanied by an abrupt increase of gate

---

**Fig. 1. Schematic description of the FET channel depletion layer profile in pinched operation.**

are made of silicon, the numerical examples are confined to this material.

**CURRENT CONDUCTION IN PINCHED OPERATION**

Extrapolation of the gradual-channel model beyond pinchoff implies that, for a drain-gate potential \( V_{de} \) in excess of the device pinchoff...
current. In physical FET structures, no such effects are observed at the onset of pinched operation.

(c) Thermal generation of carriers within the channel depletion layer. This is due to the presence of deep lying centers in the depletion layer. The magnitude of this current is on the order of the reverse current of a p-n junction diode of comparable dimensions. Thus it cannot account for a significant fraction of the observed drain current.

As indicated by the above discussion, none of the possible conduction mechanisms which may describe the current flow through a depleted layer can account for the observed behavior and magnitude of the drain current in pinched operation. Therefore, it is postulated that the channel is not totally depleted; instead, an effective residual channel filament of thickness δ, remains through the drain space-charge region (Region II). Since carrier concentrations and drift velocities are at all times finite in physical device structures, the width of this channel opening has to be finite for non-zero drain current. In the analysis, the transition between the two regions of the channel is assumed to be abrupt for mathematical compactness.

The formation of a space charge region, Region II, at the drain end of the channel can be inferred from the extrapolation of the gradual channel model to the point of pinch-off. However, this first order model cannot explain the finite length, Ld, of this region, and its dependence on the drain potential in excess of the pinch-off voltage. In order to do so, it is necessary to consider a combination of the two following effects:

(i) Failure of the gradual-channel model in the vicinity of the drain for \( V_{ds} \geq V_p \).

The gradual channel model is derived on the assumption that the longitudinal electric field gradients within the channel depletion layer are negligible compared to the transverse gradients. Consequently, the potential distribution, \( V(x, y) \), within the depletion layer can be approximated from the solution of the Poisson's equation in one dimension:

\[
\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} \approx \frac{qN(x, y)}{\varepsilon}. \tag{1}
\]

Where

- \( q = \) electronic charge
- \( N(x, y) = \) donor impurity concentration in the n-type channel
- \( \varepsilon = \) permittivity of the material.

Defining the range of validity of the gradual-channel model as that region of the channel where:

\[
\left| \frac{\partial^2 V}{\partial x^2} \right| \leq \left| \frac{\partial^2 V}{\partial y^2} \right| \tag{2}
\]

one can show that the model is valid only for that region of the channel where:

\[
W_b(x) = W_1(x) + W_2(x) \geq \frac{I_d}{A} \left( \frac{L_d \rho}{2Z V_p} \right)^{2/3}. \tag{3}
\]

\( W_b(x) = \) height of the undepleted channel
\( A = \) vertical distance between the gates
\( I_d = \) saturation value of the drain current at a given gate bias (for \( V_{gs} = V_{gd} \))
\( \rho = \) average channel resistivity
\( V_p = \) pinch-off voltage with \( V_{gs} = V_{gs0} \), and
\( x = \) depth of the channel, measured perpendicular to the x and y axes.

As indicated by equation (3), the undepleted channel height, \( W_b \), cannot go to zero for any finite drain current without violating the gradual channel approximation.

(ii) Saturation of carrier drift velocity

As indicated by the experimental data of Fig. 2\(^{10-11}\), carrier drift velocity \( v \) in silicon does not increase indefinitely with the increasing electric field, but saturates at a maximum value of \( v_0 \), as a result of lattice scattering. The rapid saturation of this carrier velocity allows one to define a critical field \( E_0 \), depending on the charge carrier type, above which velocity saturation becomes dominant. For non-degenerate materials, \( E_0 \) and \( v_0 \) are independent of the impurity doping, and depend only on the type of charged carrier.

Defining \( E_0 \) as the value of the electric field at which \( v \) attains 95 per cent of its saturation value, results in the following values of \( v_0 \) and \( E_0 \) in
For typical silicon planar FET structures, one observes that
\[ \frac{\delta_{10}}{W_m} \gg 1, \]
thus indicating that the carrier velocity saturation is the dominant effect in determining the interface between the two regions of the channel in pinched operation. Consequently, for the following analysis, only the velocity-limiting phenomenon is considered as the dominant effect in determining the dimensions of Region II. Similarly, the interface between the two regions of the channel is marked by the onset of velocity saturation, i.e.,
\[ E = E_{0}^{-\frac{1}{10}} \]

**DEVICE MODEL IN PINCHED OPERATION**

A schematic diagram of the channel depletion layer profile for an n-channel FET in pinched-mode operation is shown in Fig. 1. The p-type gate regions are assumed to be doped more heavily than the channel. The net donor concentration, \( N \) in the channel is assumed to vary in the transverse direction only, i.e.,
\[ N = N(y). \]

The two gate sections are assumed to be parallel, and independently biased at gate-source potentials of \( V_{gs1} \) and \( V_{gs2} \), respectively.

As indicated in Fig. 1, the non-uniform channel doping profile, \( N(y) \), as well as the dissimilar gate bias voltages \( V_{gs1} \) and \( V_{gs2} \) cause the channel depletion layer profile associated with each gate section to be different. Thus, the physical device structure exhibits a lack of symmetry along the transverse direction. However, following the procedure outlined in Appendix I, one can define an effective plane of symmetry through the channel, labelled as \( K-K' \). The location of this plane is independent of the source-drain potential \( V_{ds} \) in pinched operation, and is uniquely defined for a given choice of \( N(y) \), \( V_{gs1} \) and \( V_{gs2} \). The component

\[ \delta_{10} = \lim_{x \to x_0} \frac{W_{1}(x)}{qN_{w}v_{o}x}, \]

The relative significance of the two effects discussed above in determining the dimensions of the drain space charge region, Region II, can be qualitatively estimated by comparing the magnitudes of \( \delta_{10} \) and \( W_{m} \), where \( W_{m} \) is that value of \( W_{1} \) which reduces equation (3) to an equality.

\[ v_{o} \approx 10^{7} \text{ cm/sec for electrons} \]
\[ v_{o} \approx 6 \times 10^{8} \text{ cm/sec for holes} \]
\[ E_{0} \approx 2 \times 10^{4} \text{ V/cm for electrons} \]
\[ E_{0} \approx 3 \times 10^{4} \text{ V/cm for holes.} \]
of the electric field normal to the $K-K'$ plane is zero along the entire length of the channel; therefore, there is no net transfer of mobile charge across this plane. Consequently, one can separate the two-gate FET structure shown in Fig. 1 into two single-gate FETs about the $K-K'$ plane, and consider each half of the device separately.

Figure 3 shows a schematic diagram of the depletion layer profile within the active channel region* about the effective symmetry plane. The coordinate axes and the device dimensions to be used in the following analysis are also defined in this figure.

![Diagram](image)

**Fig. 3.** Depletion profile within the active channel in pinched operation, sectioned about the effective axis of symmetry.

In Region I, of length $L_1$, current conduction is described by the gradual-channel model. The channel depletion layer profile, $h(x)$, can be determined by an approximate solution of the one-dimensional Poisson equation within the channel-gate depletion layer. A detailed analysis of current flow in this region has been carried out by Bockemuehl and Richer for a generalized impurity distribution, $N(y)$, and will not be repeated here.

The generalized expression for the drain current can be written as shown in equation (4). In dealing with the one-dimensional impurity distribution, i.e., $N(x, y) = N(y)$, it is convenient to introduce a

short-hand notation for the mathematical operation of integration as follows:

$$ N(y) = \int N(y) \, dy $$

and

$$ N(y) \equiv [N(y)]. $$

Similarly, a definite integral can be expressed as

$$ \int_{a}^{b} N(y) \, dy = N(b) - N(a). $$

In terms of this notation, $N_d(x)$ is given by

$$ N_d(x) = \frac{1}{W(x)} [N(a) - N(h)]. $$

As a result of the ohmic drop along the undepleted channel, $W(x)$ is a monotonically decreasing function of $x$. Since the current, $I_d$, is continuous through the channel, the carrier drift velocity, $v(x)$, is forced to increase monotonically as a function of $x$ in order to sustain the drain current. Ultimately, $v(x)$ saturates, at which point the channel height attains its lower limit, $\delta_0$, corresponding to $v(a) = v_0$, as defined earlier in equation (5).

Formation of the residual channel of thickness $\delta_0$ as a result of velocity saturation marks the

---

*The term "active channel" is used to imply the section of the channel directly between the two gate regions, as shown in Fig. 1.
interface between the two regions of the model, since the conduction through the channel is no longer ohmic. For typical impurity profiles observed in practical device structures, and for device geometries having channel length-to-width ratios \((L_1/a)\) in excess of 5, it can be shown that\(^{(19)}\):

\[ \delta_0 \ll a. \]  

(12)

Therefore, from equation (11):

\[ N_w(L_1) = \frac{1}{\delta_0} [N(a) - N(a - \delta_0)] \approx N(a). \]  

(13)

Thus,

\[ \delta_0 \approx \frac{I_d}{qN(a) \nu_0 \sigma}. \]  

(14)

The channel-gate potential, \(V_o\), needed to extend the depletion layer a distance of \((a - \delta_0)\) into the channel, and hence cause velocity saturation, can be related to the device pinch-off voltage, \(V_p\), as:

\[ V_o = V_p - \frac{(q/\epsilon)a \delta_0 N(a)}{N(a) - N(a - \delta_0)}. \]  

(15)

where \(\epsilon\) is the permittivity of the semiconductor material. For a generalized doping profile \(N(y)\), \(V_p\) can be expressed as:

\[ V_p = (q/\epsilon)[N(a) - N(y) - \alpha N(a)]. \]  

(16)

As indicated by equation (15), due to the finite value of \(\delta_0\), pinched mode operation sets in at a value of the drain-gate potential somewhat lower than \(V_p\). However, for practical device geometries, \(\delta_0 \ll a\), therefore:

\[ V_o \approx V_p. \]  

(17)

The drain potential in the formation of the drain-gate potential somewhat lower than \(V_p\), needed for velocity saturation, \(V_d\), is defined as:

\[ V_d = V_o + V_y = V_o + V_x - V_p. \]  

(18)

This excess drain potential results in the formation of the drain in the vicinity of the drain, where the current flow is non-ohmic and is confined to an effective residual channel of thickness \(\delta\) along the \(K-K'\) plane. As will be shown later, the strong longitudinal field gradients in this region cause the free carrier concentration within the residual channel to deviate from its thermal equilibrium value of \(N(a)\), in a manner similar to that discussed by JUND and GROSVALET\(^{(13)}\).

In Region II, the potential distribution must be determined through the solution of the two-dimensional Poisson's equation with appropriate boundary conditions. For mathematical compactness, the coordinate axes are moved to the interface between the two regions of the device, as shown in Fig. 1. This is accomplished by defining a new independent variable \(x'\), where:

\[ x' = x - L_1. \]  

(19)

The potential distribution in this region must satisfy:

\[ \nabla^2 V(x', y) = \frac{\partial^2 V}{\partial (x')^2} + \frac{\partial^2 V}{\partial y^2} = -\frac{qN(y)}{\epsilon} \]  

(20)

for:

\[ 0 < x' < L_2 \]

\[ 0 < y < (a - \delta) \approx a. \]  

(21)

To ensure a continuous transition between regions I and II, we invoke the following four boundary conditions:

1. \(V(x', 0) = -V_x\).

(22)

2. \(\frac{\partial V}{\partial y}(x', a) = -E_x(x', a) = 0\).

(23)

3. \(\frac{\partial V}{\partial x}(0, a) = -E_x(0, a) = E_0\).

(24)

4. \(V(0, y) = V_o - (q/e)[yN(a) - N(y)]\).

(25)

Boundary condition 1 treats the metallurgical channel-gate junction as an equipotential plane, which is valid for a \(p^+\) gate. Condition 2 is an outcome of the effective channel symmetry about \(y = a\) (the \(K-K'\) plane). Condition 3 requires the continuity of the electric field within the conducting channel. Condition 4 ensures the continuity of the potential across the entire Region I-Region II interface. The potential distribution on right-hand side of equation (25) is that given along the \(x = L_1\) boundary of Region I by the gradual-channel approximation\(^{(2)}\).

As shown in Appendix II, the above boundary conditions are not sufficient to determine an exact solution of equation (20). However, by using a
self-consistent approach, it is possible to obtain the approximate solution:

\[
V(x', y) = -V_g + \frac{(q/e)[N(y) - N(0) - yN(a)]}{2E_Oa} + \frac{\pi y}{\pi} \sinh \frac{\pi x'}{2a} \sinh \frac{\pi x}{2a}.
\] (26)

Along the center line of the channel, this reduces to:

\[
V(x', 0) \approx V_g - V_s - \frac{2E_Oa}{\pi} \sinh \frac{\pi x'}{2a}.
\] (27)

From equation (27), the dependence of the length of the pinched region on the drain potential may be determined as:

\[
L_2 = \frac{2a}{\pi} \sinh^{-1} \left[ \frac{V_d - \pi n}{2E_Oa} \right].
\] (28)

Assuming that an abrupt boundary exists between the residual channel and the depletion layer, and that the carrier concentration within the residual channel in the direction of current flow is uniform, the excess free carrier concentration within the conducting filament of Region II can be approximated from (27), by solving the Poisson's equation along the centerline of the channel, for \(0 < x' < L_2\):

\[
\frac{n - N(a)}{N(a)} = \frac{\pi E_O}{2aqN(a)} \sinh \left( \frac{\pi x'}{2a} \right),
\] (29)

where \(n\) is the actual value of the free carrier concentration.

Since \(I_d\) is continuous throughout the channel and the carrier saturation velocity \(v_s\) is fixed, the increase of free carrier concentration as predicted by equation (29) results in a monotonic decrease of the residual channel opening, \(\delta\), as a function of \(x'\), as given below:

\[
\delta(x') = \delta_0 \left[ \frac{1}{1 + \frac{\pi E_O}{2aqN(a)} \sinh \left( \frac{\pi x'}{2a} \right)} \right].
\] (30)

Note that, since at all times \(\delta_0 \geq \delta(x')\) for \(x' \geq 0\), the basic assumptions used in obtaining the approximate solution of equation (26) are not affected by finite deviations from charge neutrality within the residual channel. However, the deviation of \(n\) from \(N(y)\) may cause the potential along the centerline of the channel to be slightly higher than that predicted by equation (27) by an amount \(\Delta V\), where

\[
\Delta V \approx \frac{q\delta^2}{2e}[n - N(a)].
\] (31)

Due to small dimensions of \(\delta\), this corrective term may be neglected.

At this point, it is instructive to examine the electric-field and the potential distribution predicted by the model, by means of a numerical example. We shall consider the case of a symmetrical \(n\)-channel silicon FET with a uniformly doped channel, with the following device parameters:

\[
N(y) = N_D = 10^{15} \text{ cm}^{-3} (5 \Omega \text{-cm, channel resistivity})
\]

\[
a = 3 \mu\text{m}
\]

\[
L = 25 \mu\text{m}
\]

\[
E_O = 20 \text{ kV/cm}
\]

\[
V_s = qN_Da^2/2e = 6.75 \text{ V}
\]

Keeping the drain voltage, \(V_d\), constant at 20 V, consider the case of two separate values of the gate bias, namely \(V_g = 0\) and 5.5 V, respectively.

For \(V_g = 0\), one calculates the following values:

\[
V_d = 13.25 \text{ V}
\]

\[
\delta_0/a = 0.036
\]

\[
L_1 = 21.3 \mu\text{m}
\]

\[
L_2 = 3.7 \mu\text{m}
\]

For \(V_g = 5.5 \text{ V}\):

\[
V_d = 18.75 \text{ V}
\]

\[
\delta_0/a = 0.001
\]

\[
L_1 = 20.5 \mu\text{m}
\]

\[
L_2 = 4.5 \mu\text{m}
\]

Note that the \(\delta_0/a = a\) assumption is well satisfied for both cases.

Figure 4 shows the electric field and the potential along the centerline of the channel \((y = a)\) for \(V_g = 0\). The predicted form of the equipotential contours through the body of the channel for zero gate bias condition is shown in Fig. 5. Note that, as a result of the boundary conditions used in the solution of the two-dimensional Poisson equation in the pinched region, the potential distribution across the \(x = L_1\) plane is continuous. In sketching the equipotential contours of Fig. 5, it is assumed that there are a sufficient number of bound charges
Fig. 4. Potential and the electric field in the conducting channel, for operation beyond pinch-off. (Values correspond to the numerical example in the text, with $V_t = 0$.)

Fig. 5. Equipotential contours within the channel for pinched operation. (Values correspond to the numerical example in the text, with $V_t = 0$.)
beyond the active drain region on which the electric field lines can terminate. These bound charges are provided by the extension of the depletion contour beyond the active channel region, as shown schematically in Fig. 1.

In terms of the device dimensions and properties discussed in the numerical example of this section, a graphical description of the variation of the free carrier concentration along the residual channel, as predicted by equation (29) is shown in Fig. 6 for the condition where \( V_s = 0 \). Using mode operation can be obtained as:

\[
I_d = \frac{zq_0^2 \mu}{\epsilon L_1} \int_{h_0}^{a-\delta_r} yN(y)[N(a-\delta_r) - N(y)] \, dy
\]  

(32)

Since \( \delta_r \ll a \), equation (26) can be written as:

\[
I_d = \frac{zq_0^2 \mu}{\epsilon L_1} \int_{h_0}^{a} yN(y)[N(a) - N(y)] \, dy
\]  

(33)

\[
\frac{\partial I_d}{\partial V_d} = -\frac{1}{L_1} \frac{\partial L_1}{\partial V_d}
\]

\[
= -\frac{V_d}{L_1} \frac{\partial L_1}{\partial V_d}
\]

\[
= \frac{I_d}{L_1} \frac{\partial L_1}{\partial V_d}
\]

**Fig. 6. Deviation from charge neutrality within the residual channel of Region II. (Values correspond to the example in the text, with \( V_s = 0 \).)**

this data, the magnitude of the corrective term, \( \Delta V \), of equation (31) is on the order of a few millivolts.

**DEVICE PARAMETERS BEYOND PINCH-OFF**

The physical model for pinched-mode operation can now be employed to determine the device parameters for operation beyond pinch-off. The particular device parameters to be examined are the drain resistance, \( r_d \), transconductance \( g_m \), and the transit time \( \tau \).

(a) **Drain resistance, \( r_d \)**

Following Bockenhul's formulation, an expression for the drain current \( I_d \) in pinched-

The only parameter in (33) showing an explicit dependence on \( V_d \) is \( L_1 \). Therefore, differentiating equation (33) with respect to \( V_d \), with \( V_s \) constant, we have:

Differentiating equation (33) with respect to \( V_d \), with \( V_s \) constant, we have:

\[
\varepsilon_d = \frac{1}{r_d} = \frac{\partial L_1}{\partial V_d} = -\frac{I_d}{L_1} \frac{\partial L_2}{\partial V_d} = \frac{I_d}{L_1} \frac{\partial L_2}{\partial V_d}
\]

(34)

Realizing that

\[
L_1 + L_2 = L_1 \quad \text{and} \quad L_1 \gg L_2
\]

(35)

equation (34) reduces to

\[
\varepsilon_d = \frac{\partial I_d}{\partial V_d} \approx \frac{I_d}{L_1} \frac{\partial L_2}{\partial V_d}
\]

(36)
Evaluating

\[ r_d = \frac{1}{E_0} \left( \frac{\partial L_d}{\partial V_d} \right) \left( \frac{\partial V_d}{\partial N_d} \right) \]

from equation (28):

\[ r_d = \left( \frac{1}{E_0} \right) \left[ \frac{L_i E_0}{I_a} \left( \frac{V_d}{2E_0 a} \right)^{1/2} \right]. \quad (37) \]

Even though \( r_d \) appears to be independent of \( V_d \), it should be remembered that the influence of the gate bias is included implicitly, through \( I_a \) and \( V_{dp} \). As indicated by the definition (18), for relatively large values of drain bias, \( V_{dp} \) shows a weak dependence on \( V_d \). Therefore, equation (37) implies that for any particular value of drain potential, the drain current resistance product, \( I_a r_d \), is relatively constant.

For operation of the device in the vicinity of pinch-off, i.e. \( V_{dp} \approx V_j \), the first term in (37) is dominant, and the dynamic drain resistance can be expressed as

\[ r_d \approx \frac{L_i E_0}{I_a}. \quad (38) \]

It should be noted that equation (38) may be used as an alternate means of determining \( E_0 \), from measured values of \( I_a \) and \( r_d \) at the point of pinch-off. The values of \( E_0 \) calculated in this manner are in very close agreement with those determined from the data of Fig. 2.

As the value of \( V_{dp} \) is increased, the second term in equation (37) becomes dominant very rapidly, so that

\[ r_d \approx \frac{\pi L_i V_{dp}}{2a L_d}. \quad (39) \]

As indicated by equation (39), the incremental drain resistance is proportional to the excess drain voltage for operation well into the pinched-mode.

It is interesting to examine the temperature dependence of the \( I_a r_d \) product for operation within the pinched region. Differentiating equation (39) with respect to temperature \( T \),

\[ \frac{\partial}{\partial T} (I_a r_d) = \frac{\pi L_i V_{dp}}{2a} \frac{\partial V_{dp}}{\partial T}. \quad (40) \]

But:

\[ \frac{\partial V_{dp}}{\partial T} = \frac{\partial}{\partial T} (V_d + V_j - V_p) \approx -\frac{\partial V_p}{\partial T}. \quad (41) \]

The temperature coefficient of the pinch-off voltage \( V_p \) is the same as that of the built-in junction potential and is of the order of \((-2 \text{ mV})/^\circ \text{C}\) for silicon. Therefore, equation (40) can be re-written as:

\[ \frac{\partial}{\partial T} (I_a r_d) \approx \frac{\pi L_i}{a} \text{ mV}/^\circ \text{C} \quad (42) \]

For most commercially available devices, the right-hand side of equation (42) is in the range of 10–100 mV C. However, the \( I_a r_d \) product beyond pinch-off [as predicted by equation (39)] is very large (typically in the 100–400 V range). Consequently, this product is relatively insensitive to temperature changes, even though its individual terms are not.

From equations (39) and (42), the percentage variation of the \( I_a r_d \) product is:

\[ \frac{\partial (I_a r_d)}{\partial T} \approx \frac{0.2}{V_{dp}} \%/^\circ \text{C}, \quad (43) \]

where \( V_{dp} \) is expressed in volts.

(b) Transconductance, \( g_m \)

The device transconductance, \( g_m \), is defined as:

\[ g_m = \left( \frac{\partial I_a}{\partial V_j} \right) \quad \text{at} \quad V_j = \text{constant}. \quad (44) \]

Using the expression for \( I_a \) as given by equation (33):

\[ g_m = -\frac{\partial I_a}{\partial V_j} = \frac{I_a}{L_i} \left( \frac{\partial L_i}{\partial V_j} \right) + \frac{L_i}{L_1} \frac{\partial E_0}{\partial V_j} \quad (45) \]

where \( g_m \) is the transconductance obtained by direct extrapolation of the gradual-channel model for operation beyond pinch-off, and is given* as:

\[ g_m = \frac{\mu_2 q}{L_1} \left[ N(a) - N(h_o) \right]. \quad (46) \]

* This ignores the dependence of low-field mobility on the impurity concentration.
However:

$$\frac{\partial L_2}{\partial V_s} = \frac{\partial L_2}{\partial V_d} \quad \text{and} \quad L_1 \simeq L_t.$$  \quad (47)

Thus, equation (45) can be approximated as:

$$g_m \simeq g_m + g_d$$  \quad (48)

where $g_d$ is the incremental channel conductance in the pinch-mode operation, as defined by equation (37).

As indicated by equation (48), the model predicts that the device transconductance $g_m$ in pinch-mode operation is greater than that predicted by the gradual-channel model by an amount equal to the drain conductance, $g_d$. For practical device structures, typical values of $g_m$ and $g_d$ are of the order of 3 m\(\Omega\) and 50 \(\mu\)\(\Omega\) respectively. Therefore, the predicted increase of $g_m$ over $g_m$ is of the order of a few per cent.

(c) Transit time

In terms of a direct extrapolation of the gradual-channel model for pinched operation, the transit time $\tau_0$ for the intrinsic device at the point of pinch-off can be expressed as:\(^{(9)}\)

$$\tau_0 = \frac{\varepsilon L_1^2}{\mu q} \mathcal{Q}(h_o),$$  \quad (49)

where $\mathcal{Q}(h_o)$ is a monotonically increasing function of $h_o$ expressed as:

$$\mathcal{Q}(h_o) = \frac{\int_{h_o}^{a} a^2 y N(y) [N(a) - N(y)]^2 \, dy}{\int_{h_o}^{a} y N(y) [N(a) - N(y)] \, dy}.$$  \quad (50)

In terms of the physical model for pinched operation developed in the previous section, the transit time $\tau$ associated with the intrinsic device may be expressed as:

$$\tau = \tau_1 + \tau_2 = \frac{\varepsilon L_1^2}{\mu q} \mathcal{Q}(h_o) + \frac{L_2}{v_o},$$  \quad (51)

where $\tau_1$ and $\tau_2$ are respectively the transit times through Regions (I) and (II) of the channel in pinched operation. Since $L_1 \gg L_2$:

$$\tau = \tau_0 - \frac{L_2}{L_t} \left[ 2\tau_0 - \frac{L_1}{v_o} \right].$$  \quad (52)

However:

$$\tau_0 \gg \frac{L_1}{2v_o}.$$  \quad (53)

Thus, an approximate expression for the transit time in the pinched-mode of operation is:

$$\tau = \tau_0 \left(1 - \frac{2L_2}{L_t} \right) \quad \text{for} \quad L_2 \ll L_t.$$  \quad (54)

Using the values of device parameters discussed in connection with the numerical example of the previous section with $V_d = 3V_p$ and $V_s = 0$, the result given by equation (54) indicates that an improvement of $\approx 20$ per cent can be obtained in $\tau$ over $\tau_0$.

EXPERIMENTAL RESULTS AND DISCUSSIONS

In order to verify the results predicted by the physical model, several commercially available silicon planar FETs of varying geometries and impurity doping profiles were used as test units. In all cases, the device geometries chosen satisfied the $L_d/a \geq 5$ condition. The values of the carrier saturation velocity, $v_o$, and the critical field, $E_o$, were computed from the independent experimental data of Fig. 2.\(^{(8-9)}\)

For purposes of device design and application, the most significant result of the model for pinched operation is the presence of a finite drain resistance which can be readily related to the device parameters as well as to the choice of the operating point. Furthermore, the model predicts that the drain "current--resistance product," $I_d R_d$, will be relatively constant, and independent of temperature for a given value of the drain potential beyond pinch-off.

Since these properties of the device show the most marked difference in device characteristics above and beyond those suggested by the gradual-channel model, they were chosen as the key parameters to be measured to verify the range of validity of the model. Detailed measurements of other device parameters are described in Ref. 9.

Experimental results are presented here for some typical test devices, whose properties are
Table 1. Properties of test units

<table>
<thead>
<tr>
<th>Device type</th>
<th>2N2499</th>
<th>2N3458</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>p-channel</td>
<td>n-channel</td>
</tr>
<tr>
<td>Gate geometry</td>
<td>double diff.</td>
<td>epitaxial</td>
</tr>
<tr>
<td>Channel length, L</td>
<td>12 (\mu)m</td>
<td>8 (\mu)m</td>
</tr>
<tr>
<td>(V_p)</td>
<td>3.8 V</td>
<td>2.5 V</td>
</tr>
<tr>
<td>(d)</td>
<td>1.2 (\mu)m</td>
<td>0.7 (\mu)m</td>
</tr>
<tr>
<td>(v_o)</td>
<td>(6 \times 10^6) cm/sec</td>
<td>(10^7) cm/sec</td>
</tr>
<tr>
<td>(E_o)</td>
<td>30 kV/cm</td>
<td>20 kV/cm</td>
</tr>
</tbody>
</table>

Figure 7 gives a graphical comparison of the measured and the calculated values of the incremental drain resistance, \(r_d\), as a function of excess drain potential in pinched operation with zero applied gate bias for each of the test devices. The theoretical values of \(r_d\) are calculated from the device geometry and properties listed in Table 1 and from equation (37) of the text. The predicted and measured values of \(r_d\) are shown in Fig. 8 as a function of \(I_o\), with constant drain bias beyond pinch-off, i.e., for

Fig. 7. Measured and calculated values of the dynamic drain resistance, \(r_d\), in pinched operation.

Fig. 8. Measured and predicted values of \(I_o\) and \(r_d\) in pinched operation, for constant \(V_{dp}^*\).
\[ V_{ds} = V_d + V_s - V_g = \text{constant} \]

Here, the theoretical curve is that given by equation (39), since \( V_{ds} > V_g \). Note that this curve shows close agreement to the measured data in both trend as well as in actual numerical values.

Figure 9 shows the measured temperature dependence of the \( I_d r_d \) product for operation in the pinched-mode with zero applied gate bias. As indicated by the observed results, both \( I_d \) and \( r_d \) vary over a wide range of values as a function of

(a)

(b)

Fig. 9. Temperature dependence of device parameters beyond pinch-off.
temperature; however their product remains relatively unchanged. This is in excellent agreement with the results predicted by equation (43).

Acknowledgments—The authors wish to thank Mr. L. Evans of Siliconix, Inc., for making available the necessary design data on a number of devices used in this program.

REFERENCES

APPENDIX I
Determination of the effective axis of symmetry
Asymmetry of the channel depletion layer profile in pinched operation comes about as a result of one or more of the following causes:
(a) Different bias voltages $V_{c1}$ and $V_{c2}$ applied to gates 1 and 2.
(b) Non-uniform (asymmetric) channel doping profile.

When either or both of these asymmetry conditions are present, the location of the effective plane of symmetry, the $K-K'$ axis of Fig. 1, can be uniquely defined for a given set of bias voltages and doping profile.

Let the respective channel depletion layer thicknesses $h_1(x)$ and $h_2(x)$ within Region 1 be related to the gate channel reverse bias, $V_{re}(x)$ at any point along the channel as given below:

$$h_1(x) = \phi_1(V_{re1}),$$

$$h_2(x) = \phi_2(V_{re2}).$$

The gate-channel voltages $V_{re1}$ and $V_{re2}$ for the respective sections of the device can be expressed as

$$V_{re1} = V_{r1} + V_r(x),$$

$$V_{re2} = V_{r2} + V_r(x).$$

where $V_r(x)$ is the potential within the conducting channel at a distance $x$ from the source, for $0 \leq x \leq L_1$. The functions $\phi_1$ and $\phi_2$ are monotonically increasing functions of $V_{re1}$ and $V_{re2}$, respectively. It will be assumed that the residual channel openings $\delta_1$ and $\delta_2$ in Region II are small compared to the respective metallurgical channel widths, i.e.:

$$\delta_1 \ll a_1 \quad \text{and} \quad \delta_2 \ll a_2$$

Setting $V_{re1}$ and $V_{re2}$ as the respective pinch-off voltages of the two single-gate devices, one can state:

$$V_{r1} = V_{re1} + V_{c1},$$

$$V_{r2} = V_{re2} + V_{c1}.$$ Where $V_{c1}$ is the potential within the conducting channel at $x = L_1$, i.e.

$$V_{c1} = V_r(L_1).$$

Considering the depletion layer thickness at $x = L_1$:

$$h_1(L_1) + h_2(L_1) = A - (\delta_1 + \delta_2) \simeq A.$$ In terms of equations (A.1) and (A.2) the result of (A.8) can be expressed as:

$$a_1 + a_2 = \phi_1(V_{re1} + V_{c1}) + \phi_2(V_{re2} + V_{c1}) = A.$$ For a given channel and gate impurity profile, the functions $\phi_1$ and $\phi_2$ relating the depletion layer thickness across the junction to the applied potential are uniquely defined. $V_{re1}$, $V_{re2}$ and the total channel width $A$ are readily measurable quantities. Therefore, the only remaining unknown, $V_{c1}$, can be determined from the solution of equation (A.9) thereby uniquely defining the $K-K'$ plane of symmetry.

Except for the case of a uniformly doped channel, the functions $\phi_1$ and $\phi_2$ cannot be expressed in a convenient closed form, but are best described graphically. Therefore, in most applications a graphical, rather than analytical, solution of (A.9) may be preferable.

APPENDIX II
An approximate solution of Poisson's equation in the pinched region
Figure 3 shows a schematic description of the coordinate axes to be used in obtaining an approximate solution of the Poisson equation within the depletion layer of Region II. In order to ensure a continuous transition from Region I to Region II of the channel, the boundary conditions of equations (22) through (25) must be satisfied.
Let the general solution of equation (20) be of the form

\[ V(x', y) = V_1(x', y) + V_2(x', y), \quad (A.10) \]

over the range

\[
0 < x' < L_2 \\
0 < y < (a - \delta) \approx a.
\]

The two components of the solution are to be chosen such that:

\[
\nabla^2 V_1(x', y) = 0, \quad (A.11) \\
\nabla^2 V_2(x', y) = -\frac{qN(y)}{\varepsilon}. \quad (A.12)
\]

The function \( V_i(x', y) \) represents the solution of Laplace's equation within the space-charge region of Region II, and is chosen to have the following functional form:

\[
V_i(x', y) = \sum_{n=1}^{\infty} A_n \sin(n\pi y) \sinh(\beta_n x'). \quad (A.13)
\]

The boundary conditions to be satisfied by \( V_i(x', y) \) are:

1. \( V_1(x', 0) = 0 \)
2. \( V_1(0, y) = 0 \)
3. \( \frac{\partial V_1}{\partial y}(x', a) = 0 \)
4. \( \frac{\partial V_1}{\partial x'}(0, a) = -E_0. \) \quad (A.14)

The first two boundary conditions are satisfied by the functional choice of \( V_1 \). The third boundary condition, in conjunction with equation (A.11), requires that:

\[
\alpha_n = \beta_n = \frac{(2n-1)\pi}{2a}. \quad (A.15)
\]

The fourth boundary condition is not sufficient to determine uniquely the unknown coefficient, \( A_n \), of equation (A.13) since this condition only states that:

\[
\sum_{n=1}^{\infty} A_n (2n-1) = \frac{2aE_0}{\pi}. \quad (A.16)
\]

This discrepancy occurs because of the fact that we have made no restrictions on the potential distribution at the drain, i.e. along the \( x' = L_2 \) plane. If this potential distribution is known, then the coefficients \( A_n \) of equation (A.13) can be uniquely defined as:

\[
A_n = \frac{F_n}{\sinh\left[\frac{(2n-1)\pi L_2}{2}\right]}, \quad (A.17)
\]

where \( F_n \) is the coefficient of the \( n^{th} \) term in the Fourier series expansion of the potential along the \( x' = L_2 \) plane.

For a physically meaningful solution, the coefficients \( A_n \) must tend to zero very rapidly as \( n \) increases. If this were not true, the hyperbolic sine terms of \( V_1 \), as given by equation (A.13) would lead to extremely high potentials and electric fields in the vicinity of the drain, resulting in an avalanche breakdown in the channel near the drain contact. Since \( V_i(L_2, y) \) is unknown, it is assumed to be of such a form as to make \( A_n \) vanishingly small for \( n \neq 1 \). This is a valid self-consistent approximation if the exact shape of the drain contact is somewhat similar to the form of resulting equipotentials, near the vicinity of the drain (see Fig. 4). Consequently, only the first term of the series given by (A.16) is retained, and the approximate expression for \( V_i(x', y) \) becomes

\[
V_i(x', y) \simeq \frac{2aE_0}{\pi} \frac{xy}{2a} \sinh \frac{\pi x'}{2a}. \quad (A.18)
\]

Let the particular solution, \( V_0(x', y) \) be described by a general expression of the form:

\[
V_0 = \sum_{k=1}^{\infty} C_k (x')^k + \sum_{j=0}^{\infty} D_j(y)^j + \varphi(y), \quad (A.19)
\]

where \( \varphi(y) \) is an arbitrary function of \( y \).

The boundary conditions to be satisfied by \( V_0(x', y) \) are:

1. \( V_0(x', 0) = -V_\varepsilon \)
2. \( \frac{\partial V_0}{\partial y}(x', a) = -E_0(x', a) = 0 \)
3. \( \frac{\partial V_0}{\partial x'}(0, y) = 0 \) \quad (A.20)
4. \( V_0(0, y) = -V_\varepsilon + \frac{q(a)(N(y) - N(0))}{(N(a) - N(0))} \)

The form of equation (A.19) which satisfies the boundary conditions of (A.20) as well as equation (A.12) is the one where:

\[
C_k = 0 \quad k = 1, 2, 3 \ldots, \quad (A.21)
\]

\[
D_j = 0 \quad j = 3, 4, 5 \ldots, \quad (A.22)
\]

\[
D_0 = -V_\varepsilon, \quad (A.23)
\]

\[
D_1 = (-q(a)(N(a) - N(0))), \quad (A.24)
\]
\[ \phi(y) = (q/e)[\overline{N(y)} - y\overline{N(0)} - \overline{N(0)}] - D_2a \quad (A.25) \]

Therefore, the total approximate solution becomes:

\[ V(x', y) = V_1 + V_2 = -V_s + (q/e)[\overline{N(y)} - \overline{N(0)} - y\overline{N(a)}] \quad (A.26) \]

\[ \times \sinh \frac{\pi x'}{2a} \]

which is the result stated in equation (26) of the text.
Generalized Proof of Shockley’s Positive Conductance Theorem

Abstract—Shockley’s theorem states that a semiconductor with a negative differential mobility and a well-behaved cathode contact, has a positive differential conductance. The proof of this theorem is generalized to arbitrary impurity distributions and geometries.

In 1954, Shockley showed\(^1\) that the static differential conductance at the terminals of a piece of semiconductor with negative differential mobility would be positive if the nature of the cathode contact is such that the electric field there is maintained near zero. In his proof Shockley assumed that the ionized impurity density is constant, and that the geometry is one-dimensional. We wish to show that Shockley’s theorem holds for arbitrary impurity distributions and geometries as long as the local mobility remains isotropic.

We first wish to retain temporarily the assumption of a one-dimensional geometry.

The electric field inside such a semiconductor obeys Poisson’s equation (using obvious sign conventions)

\[
\frac{\partial}{\partial x} F(x,j) = \frac{j}{\varepsilon(F)} - \frac{qN(x)}{\mu}
\]

(1)

where \(j\) is the electrical current density, \(\varepsilon(F)\) the electron drift velocity (like Shockley, we neglect diffusion), and \(N(x)\) the ionized donor density. If \(F(0,j)=0\), then near \(x=0\) the first term on the right side of (1) becomes very much larger than the second term, and the solution of (1) for small \(x\) becomes

\[
F(x,j) = \frac{\sqrt{2jx}}{\sqrt{\mu j}}
\]

(2)

where \(\mu\) is the low-field mobility.

Assume now that the current density through the structure is increased by an infinitesimal amount, from \(j_0\) to

\[
j = j_0 + \Delta j.
\]

(3)

This leads to an infinitesimal change of the electric field, from \(F_0\) to \(F_0 + \Delta F\).

For small \(x\), \(\Delta F\) follows from (2):

\[
\Delta F(x,j_0) = \frac{x}{\sqrt{2j_0 \mu j_0}} \cdot \Delta j > 0.
\]

(4)

The voltage increment associated with the current increment is given by

\[
\Delta U = \int_0^L \Delta F \, dx.
\]

(5)

A negative static differential conductance is equivalent to \(\Delta U < 0\). Because of (4), \(\Delta U\) can become negative only if the curve of \(\Delta F(x)\) crosses the \(x\) axis, from above! But at any point where \(\Delta F = 0\), the slope of the \(\Delta F(x)\) curve is, after (1), given by

\[
\frac{d}{dx} \Delta F(x) = \frac{1}{\varepsilon} \frac{j}{\varepsilon(F)} = \frac{-\Delta j}{\varepsilon} > 0,
\]

(6)

that is, \(\Delta F\) can cross the \(x\) axis only from below. Thus a static negative differential conductance cannot occur under the assumed conditions, irrespective of \(N(x)\).

The conclusion remains valid if \(F(0,j)\neq 0\), as long as the cathode boundary conditions are “well-behaved,” that is,

\[
\frac{\partial}{\partial j} F(0,j) > 0.
\]

(7)

---


Manuscript received June 1, 1970.
For cathode boundary conditions such that

$$\frac{\partial}{\partial j} F(0,j) < 0, \quad (8)$$

it is known that Shockley's theorem need not hold. In this case the \( \Delta F \) curve would start out below the \( x \) axis and if no crossover takes place the static differential conductance will be negative. According to (2), crossover is permitted in this case and its occurrence is necessary for a positive static differential conductance, although even with crossover the conductance may remain negative.

The neglect of diffusion is not believed to have any effect on the validity of the theorem, since diffusion effects play a role only in regions of low electric field, and there they do not reverse the sign of the current dependence of \( F(x,j) \).

In an arbitrary geometry the current density will not be independent of position, and the field lines will in general not be parallel. It is useful to consider the field lines of the incremental field \( \Delta F \). If the semiconductor is isotropic, then, since the current lines cannot cross the boundaries of the semiconductor, the incremental field lines cannot cross either. It is then convenient to set up a curvilinear coordinate system inside the structure in such a way that the local \( x \) direction is everywhere parallel (or possibly antiparallel) to the direction of the local incremental field \( \Delta F \) (Fig. 1). In such a coordinate system, \( \Delta F \) along any particular field line can be considered a scalar \( \Delta f \), and the incremental voltage \( \Delta U \) remains as given by the scalar equation (5), if the integration is carried out along any one of the field lines, and if \( L \) is the length of the particular line chosen.

We will assume temporarily that as a result of the postulated increase in overall device current, the magnitude of the current density increases everywhere in the device. In this case (4) remains valid, and therefore \( \Delta F \) must again change its sign if the static differential conductance is to be negative. This crossover must take place along every field line, so there must be a complete "crossover surface" cutting through the device, separating the cathode from the anode region. Everywhere on this surface the total field is equal to the original field, the incremental current density \( \Delta j \) has the direction of the original current density \( j_0 \), and the change in the magnitude of the current density can still be expressed by the scalar equation (3).

Now let \( x_0 \) be the crossover point for some incremental field line, that is, \( \Delta F(x_0) = 0 \). We can again evaluate the slope of \( \Delta F \), at \( x = x_0 \), along this field line, from a three-dimensional generalization of Poisson's equation (6). To this end we consider not a single field line, but an infinitesimal "bundle" of adjacent field lines. Let \( A(x) \) be the (infinitesimal) cross-sectional area of this bundle, which in general will vary along the length of the bundle. Then we can write the three-dimensional equivalent of the one-dimensional Poisson equation (6), at \( x = x_0 \), in the form

$$\frac{\partial}{\partial x} \left[ A(x) \cdot \Delta F(x) \right] \bigg|_{x_0} = A(x_0) \cdot \frac{\partial}{\partial x} \left[ \frac{\Delta F(x)}{\varepsilon} \right] \bigg|_{x_0} = A(x_0) \cdot \frac{1}{\varepsilon} \frac{\Delta j}{\Delta F} > 0. \quad (9)$$

All other terms vanish. Clearly, \( A(x_0) \) cancels out, and all our earlier conclusions remain valid.

However, in a threedimensional geometry there is no a priori assurance that \( \Delta j \) will be positive everywhere when the total current increases. The only statement that can be made is the following. On any surface that separates cathode and anode, the current density through this surface cannot decrease everywhere, but it must increase over at least part of this surface. Now one such surface is the cathode surface itself. From the two facts that it is an equipotential surface, and that everywhere

$$\text{curl } \Delta F = 0 \quad (10)$$

it can be shown that \( \Delta j(0) \) must have the same sign across the entire cathode surface. And since it must be positive over at least part of the cathode, it must be positive over the entire cathode. Thus the conclusion remains valid that \( \Delta j \) must change its sign along every field line for negative conductance to occur. Consider now the crossover surface. Over at least part of this surface \( \Delta j > 0 \), and at least over those parts of the surface the required crossover cannot in fact take place. Therefore, the static differential conductance cannot be negative for well-behaved boundary conditions, regardless of both doping distributions and device geometry.

A negative static (not merely time-average) differential conductance is occasionally observed in experimental Gunn effect devices. Our proof dispenses with any notion that such behavior can be explained by doping nonuniformities or geometry effects, rather than imperfect cathode boundary conditions.

H. Kroener
Dep. of Elect. Eng.
University of Colorado
Boulder, Colo. 80302

Device Analysis [Arbitrary Doping, 1]

\[ Q(h) = \int_0^h \rho(y) \, dy \, (C \, cm^{-2}) \]

Abrupt Depletion Region \( \Rightarrow E_y(y = h) = 0 \)

\[ E_y = -\frac{\partial V}{\partial y} = \frac{1}{\varepsilon_s} \int_0^y \rho(y) \, dy + c \]

\[ \Rightarrow \frac{\partial V}{\partial y} = \frac{1}{\varepsilon_s} [Q(h) - Q(y)] \]

Integrating Once More

\[ V(h) = \frac{1}{\varepsilon_s} \int_0^h y \rho(y) \, dy \]

Net Voltage is \( \alpha \) the First Moment of Charge Distribution

\[ V_p = \frac{1}{\varepsilon_s} \int_0^a y \rho(y) \, dy \]

\[ \frac{dV}{dh} = \frac{h \rho(h)}{\varepsilon_s} \]

\[ C = \frac{dQ(h)}{dV} = \left( \frac{dQ}{dh} \right) \left( \frac{dh}{dV} \right) = \frac{\varepsilon_s}{h} \]

\( \Rightarrow \) Parallel Plate Capacitor Independent of Doping
Device Analysis 'Arbitrary Doping, 2')

\[ I_D = Z\mu \frac{dV}{dx} \int \rho(y)dy \]

\[ I_D dx = Z\mu \left( \frac{dV}{dh} \right) dh \int \rho(y)dy \]

\[ h = y_1 \text{ at } x = 0 : h = y_2 \text{ at } x = L \]

\[ \int I_D dx = \frac{2Z\mu}{\varepsilon_o} \int o_{h_0}^{L} h \rho(h) dh \int \rho(y)dy \]

\[ I_D = \frac{2Z\mu}{\varepsilon_o} \int_{y_1}^{y_2} \left[ Q(a) - Q(h) \right] h \rho(h) dh \]

\[ g_m \equiv \frac{\partial I_D}{\partial V_G} = \frac{\partial I_D}{\partial y_1} \frac{\partial y_1}{\partial V_G} + \frac{\partial I_D}{\partial y_2} \frac{\partial y_2}{\partial V_G} \]

\[ g_m = \frac{2Z\mu}{L} \left[ Q(y_2) - Q(y_1) \right] \]

Recall: \[ \frac{dV}{dh} = \frac{h \rho(h)}{\varepsilon_o} \]
Velocity Saturation

- A Region of Length $L_1$ is Unsaturated
- A Region of Length $L_2$ is Saturated
Device Analysis (Partial Velocity Saturation, 1)

Region 1: Gradual Channel Approximation

\[ u_c = \frac{V_c}{a} \]

\[ I_D = I_1 \left[ 3(u_c^2 - u_1^2) - 2(u_c^3 - u_1^3) \right] \]

Where

\[ I_1 = \frac{Z \mu q^2 N_D^2 a^3}{6 \varepsilon_s L_1} \]

Region 2: Velocity Saturation

\[ I_D = q N_D V_s (a - y_d) Z \]

\[ = I_s (1 - u_c) \]

\[ I_s = q N_D V_s a Z \]

\[ I_D(\text{Region 1}) = I_D(\text{Region 2}) \]

\[ \Rightarrow L_1 = \frac{z L (u_c^2 - u_1^2) - \frac{2}{3}(u_c^3 - u_1^3)}{1 - u_c} \]

\[ z = \frac{\mu V_p}{L} \cdot \frac{1}{V_s} \] Velocity Saturation Factor
Device Analysis (Partial Velocity Saturation, 2)

Region 1:

\[ V(1) = V_p - V_s = V_p (u_c^2 - u_1^2) \]

Region 2:

Longitudinal Field is Determined by Free Charges on Drain Electrode

Laplace's Equation Yields

\[ V(2) = \frac{2a}{\pi} E_c \cos \frac{\pi y}{2a} \sinh \left[ \frac{\pi (L - L_1)}{2a} \right] \]

\[ V_D = V(1) + V(2) \]

\[ V_D = V_p \left( (u_c^2 - u_1^2) + \frac{2a}{\pi zL} \sinh \left[ \frac{\pi (L - L_1)}{2a} \right] \right) \]

Recall: \( L_1 = zL \frac{(u_c^2 - u_1^2) - \frac{2}{3}(u_c^3 - u_1^3)}{1 - u_c} \)

\[ I_D = I_s (1 - u_c) \]

Recall: \( z = \frac{\mu V_p}{L} \cdot \frac{1}{V_s} \)

\[ I_s = q N_D V_s a Z \]
Extent of Velocity Saturation

![Graphs showing the extent of velocity saturation.](image)

Beyond Saturation A Large Fraction of the Channel is Velocity Saturated
Saturated Velocity Model

Complete Velocity Saturation in the Channel

\[ L_1 = 0 \quad L_2 = L \]

Uniformly Doped:

\[ I = q V_s Z (a-h) N_D \]

Arbitrary Doping:

\[ I = V_s Z \int_{a}^{h} \rho (y) \, dy \]

Recall:

\[ \frac{dV}{dy} = \frac{1}{\varepsilon_s} \int_{0}^{y} \rho (y) \, dy \]

\[ V(h) = \frac{1}{\varepsilon_s} \int_{0}^{h} y \rho (y) \, dy \]

\[ g_m = \frac{dI}{dV} = V_s \cdot Z \cdot \frac{\varepsilon_s}{h (V_0)} \]

\[ g_m = V_s \cdot C_g (F) \]
\[ I_{D} \left( V_{D}^{\text{sat}} \right) = qN_{D}V_{S} \left( a - y_{c} \right)Z \]

\[ I_{D} \left( V_{D}^{\text{sat}} + \Delta V_{D} \right) = qN_{D}V_{S} \left( a - y_{c} + \Delta y \right)Z \]

\[ - \Delta I_{D} \left( \Delta V_{D} \right) = qN_{D}V_{S} \Delta y Z \]

**IN A SQUARE LAW FORMULATION**

\[ \beta = \frac{2\varepsilon_{S}V_{S}W}{a \left[ V_{p} + 3E_{c}L_{1} \right]} \]

\[ I_{ds} = I_{DSS} \left( 1 + \lambda V_{ds} \right) \tanh \left( \eta V_{ds} \right) \]

\[ \eta \text{ assures that as } V_{DS} \rightarrow 0 \]

**MODEL - SHOCKLEY MODEL**

\[ \eta = \frac{g_{d}}{I_{DSS}} : g_{d} = g_{do} \left( 1 - u_{1} \right) g_{do} = \frac{qN_{d}\mu_{a}Z}{L} \]
Device Analysis "Uniform Doping, 2")

\[ I_{DSS} = \frac{Z\mu q^2 N_D a^3}{6\varepsilon_S L} \]

\[ u = \frac{h}{a} = \left[ \frac{V + V_G + V_{bi}}{V_p} \right]^{1/2} \]

\[ u_1 = \frac{y_1}{a} = \left[ \frac{V_G + V_{bi}}{V_p} \right]^{1/2} \]

\[ u_2 = \frac{y_2}{a} = \left[ \frac{V_D + V_G + V_{bi}}{V_p} \right]^{1/2} \]

\[ I_D = I_{DSS} \left[ 3 \left( u_2^2 - u_1^2 \right) - 2 \left( u_2^3 - u_1^3 \right) \right] \]

\[ I_{Dsat} (u_2 = 1) = I_{DSS} \left( 1 - 3u_1^2 + 2u_1^3 \right) \]

\[ = I_{DSS} \left[ 1 - 3 \left( \frac{V_G + V_{bi}}{V_p} \right) + 2 \left( \frac{V_G + V_{bi}}{V_p} \right)^{3/2} \right] \]

DEFINE \( V_T = V_{bi} - V_p \)

TAYLOR EXPANSION AROUND \( V_G = V_T \)

\[ I_D = \frac{Z\mu\varepsilon_s}{2aL} \left( V_G - V_T \right)^2 \]

SQUARE LAW
Device Analysis [Uniform Doping, 1]

\[ I(A) = q(C) \cdot n_s \left( \text{cm}^{-2} \right) \cdot V_\theta \left( \text{cm s}^{-1} \right) Z \]
\[ = q \cdot N_D \left[ a - h(x) \right] \cdot \mu \frac{d V[x]}{dx} \cdot Z \]

- Abrupt Depletion Region
- Constant Mobility
- Gradual Channel Approximation (Longitudinal Potential is Slowly Varying)

\[ h(x) = \sqrt{2 \varepsilon_s \left[ V(x) + V_G + V_{bi} / q N_D \right]} \]
\[ dV = \frac{q N_D}{\varepsilon_s} \cdot hdh \]

\[ \int_0^L I_D \, dx = I_D \cdot L = \int_{y_1}^{y_2} \frac{Z \mu q \cdot N_D (a - h) \cdot q N_D dh}{\varepsilon_s} \]
\[ I_D = \frac{Z \mu q^2 N_D^2 a^3}{6 \varepsilon_s L} \left[ \frac{3 (y_2^2 - y_1^2)}{a^2} - \frac{2}{3} \left( \frac{y_2^3 - y_1^3}{a^3} \right) \right] \]
Lateral vs. Vertical Transistors

Lateral

- Ease of Layout
- Ease of Processing
- Low Parasitic Capacitances
- Ease of Integration
- Relatively Planar

Need Sub-0.25μm Lithography

Vertical

- High Density
- Critical Dimensions
- Highly Uniform

Difficult Materials Technology
Difficult Processing Technology
High Parasitic Capacitances

High Performance

Heterostructure Devices

HEMT Like Devices
HBT, RTT
Compound Semiconductor Based Transistor Today [Industry]

UNIPOLAR TRANSISTORS

LATERAL TRANSISTORS

MESFET - GaAs (Workhorse, Hughes)
JFET - GaAs (McDonell Douglas)
  - GaInAs (LEP)
  - InP (NRL)
HEMT - AlGaAs / GaAs (GE, TRW, Toshiba, Sony, Fujitsu)
  - AlGaAs / GaInAs (GE, Honeywell, TI)
  - Al$_{48}$In$_{52}$As / Ga$_{47}$In$_{53}$As (Hughes, GE)
  - Al$_{48}$In$_{52}$As / Ga$_{47-u}$In$_{53+u}$As (Hughes, AT&T)
SISFET - AlGaAs / GaAs (IBM, NTT)
MISFET - GaAs (RTI, IBM, Honeywell)
  - GaInAs (RCA)
  - InP (NOSC)

VERTICAL TRANSISTORS

Vertical FET - GaAs (Westinghouse, Hughes)
PBT - GaAs (Lincoln Labs)
HBT - AlGaAs / GaAs (Rockwell, NTT, NEC, TI)
  - InP / GaInAs (Bellcore, AT&T)
  - AlInAs / GaInAs (Hughes, NTT)
HET - AlGaAs/GaAs (IBM, Fujitsu)
  - AllInAs / GaInAs (IBM, Fujitsu)
RTT - AlGaAs / GaAs (AT&T, Fujitsu)
  - AllInAs / GaInAs (AT&T, Fujitsu)
FET Output Conductance (Substrate Conduction)

Debye Tail From Contacts

\[ n_{\text{min}} \approx \frac{2\pi^2 \varepsilon k_B T}{q^2 L^2} \begin{bmatrix} \text{Thermal} \\ V_{DS} = 0 \end{bmatrix} \]

\[ L \ (0.5 \ \mu m) : n_{\text{min}} = 1.5 \times 10^{15} \ \text{cm}^{-3} \]

For \( V_{DS} > 0 \)

\[ \frac{\partial \varepsilon}{\partial x} = -\frac{q n(x)}{\varepsilon} \]

\[ J = q n(x) V(x) \]

For Short Channel Devices \( V(x) = V_s \)

Boundary Conditions:

\[ E(0) = 0 \]

\[ -\int_0^L E \, dx = V_{ds} \]

\[ J_{sub} = \frac{2 \varepsilon V_s V_{ds}}{L^2} \]
FET STRUCTURE AND EQUIVALENT CIRCUIT
EQUIVALENT CIRCUIT MODELS

Advantages

- Relatively simple (minimum number of elements)
- Relatively easy to formulate
- Efficient execution time
- Element values determined from terminal measurements or simple physical models

Disadvantages

- Difficult to define unique parameter values
- Elements difficult to relate to device design (simple physical models are not quantitatively accurate)
- Nonlinearities based upon dc characteristics - accurate extrapolation to RF not established
- Harmonic, intermodulation and mixing performance difficult to accurately simulate
MESFET DRAIN CURRENT MODEL

\[ I_{ds} = I_p \left| 1 - \frac{\left( V_{gs} - V_{bi} \right)}{V_p} \right| = \beta \left( V_{gs} - V_T \right)^2 \]

**SPICE 2 - Shichman and Hodges (1968)**

linear region: \( I_{ds} = \beta V_{ds} \left[ 2 \left( V_{gs} - V_p \right) - V_{ds} \right] \left( 1 + \lambda V_{ds} \right) \)

saturation region: \( I_{ds} = \beta \left( V_{gs} - V_p \right)^2 \left( 1 + \lambda V_{ds} \right) \)


\[ I_{ds} = \beta \left( V_{gs} - V_p \right)^2 \left( 1 + \lambda V_{ds} \right) \tanh \left( \alpha V_{ds} \right) \]

\[ I_{ds} = \left( A_0 + A_1 V_1 + A_2 V_1^2 + A_3 V_1^3 \right) \tanh \left( \alpha V_{out} \right) \]

Curtice includes a time delay: \( I_{ds} \left[ V_{gs} \left( t - \tau \right) V_{ds} \right] \)

\[ V_1 = V_{in} \left( t - \tau \right) \left[ 1 + \beta \left( V_{ds} - V_{out} \right) \right] \]

**Statz, Smith, Pucel and Haus (1987)**

\[ I_{ds} = \frac{\beta \left( V_{gs} - V_T \right)^2}{1 + b \left( V_{gs} - V_T \right)} \left( 1 + \lambda V_{ds} \right) \tanh \left( \alpha V_{ds} \right) \]

59
LARGE SIGNAL GaAs MESFET EQUIVALENT CIRCUIT

NONLINEAR CIRCUIT ELEMENTS

- Idg(Vout-Vin): Drain-gate voltage-controlled current source due to drain-gate avalanche breakdown
- Igs(Vin): Gate voltage-controlled current source due to forward biasing of the gate
- Ids(Vin,Vout): Drain-source voltage-controlled current source
- Cdg(Vout-Vin): Drain-gate capacitance
- Cgs(Vin,Vout): Gate-source capacitance
- Cds(Vout): Drain-source capacitance
- Rin(Vin,Vout): Gate-source charging resistance
- Rds(Vin,Vout): Drain-source resistance
PHYSICAL DEVICE MODELS

Advantages

- Accurately represent physical operation
- Physical effects easily added (e.g., breakdown, substrate conduction, surface effects, etc.)
- RF performance easy to relate to device design
- Accurate harmonic, intermodulation and mixing performance

Disadvantages

- More difficult to formulate
- Lengthy execution time (numerical models)
- RF operation not easily simulated (numerical models)
- Difficult to implement in IC environments
Power FET

DEFINITIONS

DRAIN EFFICIENCY = \frac{\text{AC Output Power}}{\text{DC Input Power}}

\eta = \frac{P_{\text{out}}}{P_{\text{DC}}} \times 100\%

Power Added Efficiency =

\frac{\text{AC Output Power} - \text{AC Input Power}}{\text{DC Input Power}}

\eta_{\text{add}} = \frac{P_{\text{out}} - P_{\text{in}}}{P_{\text{DC}}}

P_{\text{out}} (\text{Class A}) = \frac{1}{2} (\Delta I \cdot \Delta V) = \frac{1}{8} I_{fe} \left( V_A - V_s \right)

P_{\text{out}} \approx \frac{1}{8} I_{fe} V_A
Power FET (Trade-Offs, 1)

CLASS OF AMPLIFIER

Output Power vs. Efficiency

Class A  Max Power
Minimum Efficiency (25%)

Class C  Minimum Power
Maximum Efficiency

DOPING DENSITY

\[ Q_d = \int N_d(y)dy \]

As \( Q_d \rightarrow I_{fe} \)

But \( V_A = \frac{2\varepsilon F_o^2}{qA_iQ_d} \)

Where \( F_o \) = Characteristic Field for Impact Ionization and \( A_i \) is given by \( \alpha = A_i \exp \left( -\frac{F_o}{F} \right)^2 \)

\[ V_A = \frac{K}{Q_d} : K = 5.3 \times 10^{13} \text{ Vcm}^{-2} \text{ (GaAs)} \]

\[ \frac{q}{V_3KZ} \]

\[ P_{out} = \frac{8}{\text{for } V_S = 1 \times 10^7 \text{cm/s}} \]

\( P_{out} = 1 \text{W/mm} \)
TO MAXIMIZE POWER OUT, Z i

\[ V = V_0 e^{-\alpha z} \left( \cos \beta z + jsin\beta z \right) \]

\[ P_{in} = \frac{V_0^2}{r_{in}} e^{-2\alpha z} \cos^2 \beta z \]

\[ P_{out} = \frac{gV_0^2}{r_{in}} \int_0^Z e^{-2\alpha z} \cos^2 \beta z \]

\( g = \text{INTRINSIC GAIN WITH NO LOSS} \)

\[ P_{out} = \frac{gV_0^2}{r_{in}} \]

- GaAs FET PRINCIPLES & TECHNOLOGY (DILORENZO & KHANDELWAL)
Low Noise Amplification

\[ i_{nd}^2 = \text{Short-Circuit Channel Noise} = 4kT_0 \Delta f \cdot g_m \rho \]

\[ i_{ng}^2 = \text{Induced Noise Current in the Gate} = 4kT_0 \Delta f \frac{\omega^2 C_{gs}^2 R}{g_m} \]
Low Noise Amplification, 2

Sources of Channel Noise $i_{nd}^2$

- $V_{DS} = 0 \quad i_{nd}^2 \rightarrow 4kT_0 \Delta f g_m \cdot \left( \frac{G_{ds}}{g_m} \right)$
- $V_{DS} > 0$
  - Thermal Noise Voltage Gets Amplified by Changes in Conduction Channel
  - Hot Electron Noise as Electron Temperature Rises
  - Intervalley Scattering Noise
  - High Field Diffusion Noise
  - Avalanche Noise

Source of Induced Gate Noise

Noise Voltages in the Channel Cause Change in Depletion Width $\Rightarrow$ Induce $\Delta Q$ on Gate Electrode

$$i_{nG}^2 = 4kT_0 \Delta f \omega^2 C_{gs} \cdot R \frac{\omega^2 C_{gs}^2}{g_m} \cdot R$$

$V_{DS} \rightarrow 0 \Rightarrow R \sim g_m R_i$
Low Noise Amplification, 3

Noise Figure \( (F) = \frac{\text{Total Output Noise Power}}{\text{Total Output Noise Power Generated by Source}} \)

[Fukui's Equation] \( F_{\text{min}} = 1 + k_T f \frac{1}{f_T} \sqrt{g_m(R_g + R_s)} \)

\[
F = F_{\text{min}} + \frac{R_n}{R_{ss}} \left[ \frac{(R_{ss} - R_{op})^2 + (X_{ss} - X_{op})^2}{R_{op}^2 + X_{op}^2} \right]
\]

\( R_n = \frac{K_2}{g_m^2} = \text{Equivalent Noise Resistance} \)

\( R_n = 1 \Rightarrow \text{Broadband Performance} \)

\( R_{op} = K_3 \left[ \frac{1}{4 g_m} + R_g + R_s \right] \)

\( X_{op} = \frac{K_4}{f C g s} \)
Logic Schemes, 1

- $V_{out (HI)} < \Phi_n$
- Small Voltage Swing and Noise Margin
- Enhancement and Depletion Mode Devices Required
- Tight Threshold Voltage Control
- High Speed, Low Power, High Density

Direct Coupled FET Logic (DCFL)

- Requires Two Bias Voltages
- Inverter Includes Logic Branch and Driver / Voltage Shifter Branch
- Large Power Consumption
- Low Density

Low Fan-Out Sensitivity
Single Threshold Logic

Buffered FET Logic (BFL)
Logic Schemes, 2

- Capacitor Provides Isolation Between Input and Output Devices
- Require Series of Diodes to Charge the Capacitor
- FET Differential Amplifier with Buffer Stages
- Compatible with Bipolar ECL
- Threshold Voltage Requirements Not Stringent
- Input Level for Switching $\sim V_{\text{IN,1}} - V_{\text{IN,2}}$
- High Power Dissipation
Physics of Modulation Doped Heterostructure

- Electrons Are Confined Near the Heterojunction Interface Because of Coulombic Attraction of Donor Ions

- Width of Potential Well < De Broglie (Wavelength ~ 250 Å)

- Two Dimensional Quantized Electron Gas at Interface

  \[ F(x,y,z) = \Phi(z) \exp \left( \mathbf{q} \cdot \mathbf{r} \right) \]

  \[ q = 2D \text{ wave vector} \quad \mathbf{r} = 2D \text{ vector} \]

Charge and Potential Determined Self Consistently Schrödinger's Equation:

\[ \frac{\hbar^2}{2m} \frac{d^2}{dz^2} \Phi_i + \left[ E_i - V(z) \right] \Phi_i = 0 \]

Poisson's Equation:

\[ \frac{d^2 V}{dz^2} = \frac{q \rho(z)}{\varepsilon} \]

\[ \rho(z) = -q \sum_{i=0}^{\infty} n_i |\Phi_i|^2 (z) \quad (N_D = N_A = 0) \]

\[ n_i = \frac{m \cdot k_B T}{\hbar^2} \ln \left( 1 + \exp\left[ q (E_F - E_i) / k_B T \right] \right) \]
up of a drop across the oxide and a drop across the space-charge region in the silicon. Equations for these two voltages are derived in Section 8.3.

This qualitative discussion of surface charge, voltage, and fields has served to introduce the important effect of the voltage applied to the overlying metal in determining the properties of the silicon surface. Although the system is basically only a capacitor, the various forms that the surface charge in the silicon can assume cause very significant differences in the electrical properties of the silicon surface. For example, the surface can be highly conducting and electrically connected to the bulk when it is accumulated; it can be highly insulating when it is depleted of free carriers, or else it can be highly conducting, but disconnected from the bulk, when it is inverted. These three conditions can be controlled by the bias applied between the metal and the bulk of the silicon. Because of this control, the metal layer is usually called the gate, and the voltage on the metal is denoted $V_g$.

**Silicon Gate.** For many device applications, it is desirable to construct the gate from heavily doped silicon rather than from a metal (Problem 8.2). The silicon may be conveniently deposited over the gate oxide by chemical vapor deposition (CVD) as discussed in Chapter 2. Because the silicon layer is stable at high temperatures, dopant atoms can be diffused into the substrate after the gate has been deposited, and the whole system can be passivated with an overcoating oxide. The advantages of this technology become more apparent in the discussion of MOS integrated circuits in Chapters 9 and 10. Because the silicon is deposited over amorphous silicon dioxide, it is a polycrystalline film typically consisting of submicrometer-sized crystallites. Although the gate is silicon, its electrical function in the MOS system is similar to that of the metal gate, and silicon-gated structures are usually described as MOS systems.

### 8.2 Capacitance of the MOS System

As with the Schottky barrier and the $pn$ junction, analysis of the behavior of the small-signal capacitance measured between the two output electrodes provides valuable insight into the electrical behavior of the MOS system. In the case of the MOS system, this analysis has been central to research that has resulted in the present well-developed state of understanding of the oxide-silicon system and its technology. The qualitative discussion of the previous section provides a valuable framework on which to build an understanding of the behavior of the small-signal capacitance as the applied gate voltage is changed.

Consider first that the MOS system is biased with a steady voltage that causes the silicon surface to be accumulated. For a $p$-type silicon sample such as that of Figure 8.2, this corresponds to a negative applied voltage and would result in a charge configuration like that sketched in Figure 8.4. The excess holes at the surface are pulled very close to the oxide. If a small ac voltage $V_a$ is superposed on the dc bias $V_g$, it causes small variations in the charges stored on the metal gate and at the silicon surface. If $V_a$ measures the small-signal capacitance measured will be close to the modulated charge in the oxide. The more the surface is ac effectively, the Debye length and hence, the capacitance will increase with the pure oxide. Thus, approaches

$$C = \frac{L_D}{x_d}\text{.}$$

where $x_d$ is the oxide thickness of its flat-band value, the capacitance decreases as the Debye length increases. A formulation for capacitance equation under the condition contribute to the total space-charge region of the oxide capacitance and $L_D$ the extrinsic Deb.
8.2 Capacitance of the MOS System

and at the silicon surface. If the system is now connected to an instrument that measures the small-signal capacitance associated with these variations, the capacitance measured will be close to that of the oxide itself because the spatial extent of the modulated charge in the silicon is small compared to the oxide thickness. The more the surface is accumulated, the thinner will be the accumulation layer (effectively, the Debye length at the surface will be reduced by the added carriers); hence, the capacitance will asymptotically approach the capacitance associated with the pure oxide. Thus, the capacitance per unit area is C in accumulation approaches

\[ C_{ox} = \frac{e_{ox}}{x_{ox}} \]  

(8.2.1)

where \( x_{ox} \) is the oxide thickness. When the gate voltage is changed in the direction of its flat-band value, the surface accumulation decreases to zero and the capacitance decreases as the Debye length at the surface increases. To obtain an exact formulation for capacitance in this bias range, it is necessary to solve Poisson's equation under the condition that free electrons, free holes, and dopant atoms all contribute to the total space charge at the surface. This analysis was first carried out by Kingston and Neustadtter, and their results have been widely applied to obtain various electrical properties of MOS systems. A particular result that is readily calculated from the Kingston-Neustadtter theory is the capacitance per unit area when \( V_G = V_{FB} \). If this quantity is called \( C_{FB} \), it can be expressed as

\[ C_{FB} = \frac{1}{\frac{x_{ox}}{e_{ox}} + \frac{kT}{q^2\epsilon_s N_a} + \frac{1}{1/C_{ox} + L_d/e_s}} \] 

(8.2.2)

where \( L_d \), the extrinsic Debye length has been defined in Equation 4.2.14.

When the gate voltage becomes more positive than the flat-band voltage, holes are repelled from the surface of the silicon and the system is in depletion. Under this condition, relatively straightforward electrostatic analysis shows that the overall capacitance \( C \) corresponds to the capacitance obtained by a series connection of the oxide capacitance and the capacitance \( C_s \) across the surface depletion region (Problem 8.3):

\[ C = \frac{1}{1/C_{ox} + 1/C_s} = \frac{1}{1/C_{ox} + x_d/e_s} \]  

(8.2.3)

where \( x_d \) is the width of the surface depletion layer, which depends upon gate bias as well as the doping and oxide properties. From Equation 8.2.3, we see that the capacitance of the system decreases as the depletion region widens (Figure 8.7).

When the gate bias is increased sufficiently to invert the surface, a new feature must be considered to describe the MOS capacitance behavior. We recall that the
inversion layer at the MOS surface results from the generation of minority carriers. Hence, the population of the inversion layer can change only as fast as carriers can be generated within the depletion region near the surface. This limitation causes the measured capacitance to be a function of the frequency of the ac signal used to measure the small-signal capacitance of the system.

The simplest case arises when both the dc gate-bias voltage and the small-signal measuring voltage are changed very slowly so that the silicon can always approach equilibrium. In this case, the signal frequency is low enough so that the inversion-layer population can "follow" it. The capacitance of the MOS system is just that associated with charge storage on either side of the oxide; its value is therefore approximately $C_{\text{ox}}$. Under these conditions a plot of measured capacitance versus gate bias follows the dashed curve marked "low frequency" in Figure 8.7: going from $C_{\text{ox}}$ in the accumulation region of bias through a decreasing region as the surface traverses the depletion region and moving back up to $C_{\text{ox}}$ when the surface becomes inverted.

The results of Problem 8.6 show that a characteristic time to form an inversion layer at the surface of an MOS system biased to inversion is of the order of $(2N_v \tau_0 / n)$ where $\tau_0$ is the minority-carrier lifetime at the surface.* For typical values of lifetime (1 $\mu$s) and dopant concentrations ($10^{15} \text{ cm}^{-3}$), this time is roughly 0.2 s. Therefore, the small-signal measuring voltage must be changed very slowly to observe the low-frequency C-V curve (dashed curve of Figure 8.7).

Figure 8.7  (a) Small-signal capacitance of an MOS system with $p$-type silicon. Low-frequency behavior: both the bias voltage and the ac measuring signal vary slowly (less than ~10 Hz). High-frequency behavior: bias voltage varies slowly, but the ac measuring signal varies rapidly (typical circuits use 1 MHz). Deep-depletion behavior: both the gate bias voltage and the ac measuring signal vary rapidly. The behavior at point (i) is described in the text. (b) The equivalent circuit for the overall capacitance $C$ is the series connection of $C_{\text{ox}}$ (fixed) and $C_s$ (variable) with voltages $V_g$, $\phi_s$, and $V_b$ at the gate, interface, and substrate, respectively.

* Generation from surface states is considered negligible (discussed further in Section 8.5) Note that $\tau_0$ can vary considerably in practice.
The presence of some means, such as surface illumination, that can stimulate the surface generation rate will increase the range of the low-frequency behavior. If the inversion layer is able to make ohmic contact to a region with which it can exchange electrons, the low-frequency curve can be observed at frequencies extending into the MHz range. This occurs because the electrons in the inversion layer can then be supplied and withdrawn rapidly through the ohmic connection.

When the ac measuring signal is changed rapidly while the dc bias voltage is varied slowly, the inversion layer cannot respond to the measuring signal. The number of charges in the silicon space-charge layer is modulated instead by the movement of holes at the far edge of the depletion region. The capacitance then corresponds to the series combination of the oxide capacitance and the depletion-region capacitance, as was true under depletion bias. Since the depletion region reaches a maximum width \(x_{\text{dmax}}\) when the system goes into strong inversion, the measured capacitance approaches a value corresponding to the series connection of the oxide capacitance and the capacitance associated with the maximum depletion-region width. It remains constant at this value as the bias voltage is increased further. This high-frequency C-V curve is shown by the solid line in Figure 8.7.

A final capacitance behavior is sketched in Figure 8.7. This is the curve shown dot-dash and marked deep depletion. It corresponds to the experimental situation in which both the gate bias voltage \(V_g\) and the small-signal measuring voltage vary at a faster rate than can be accommodated by generation in the surface depletion region. Since the inversion layer cannot form, the depletion region becomes wider than \(x_{\text{dmax}}\) and the name deep depletion is properly descriptive. The capacitance in this mode is given by Equation 8.2.3 as was true in the normal depletion mode, here, however, \(x_d\) exceeds the inversion value \(x_{\text{dmax}}\) and \(C\) does not reach a minimum. One means of generating a “deep depletion” capacitance curve is to sweep the gate bias \(V_g\) with a low-frequency triangular wave on which is superposed the sinusoidal ac measuring frequency. The generation rate of carriers increases as the depletion layer is widened, however, and the deep-depletion curve is frequently observed to relax to the high-frequency curve at higher biases. The relaxation is indicated schematically at point (i) on the deep-depletion curve of Figure 8.7.

## 8.3 MOS Electronics

The previous two sections have given a qualitative introduction to the charge induced at the silicon surface by the voltage applied between the overlying gate and the substrate. The particular application of this theory to the MOSFET in the next two chapters demands a more quantitative background, especially for conditions under inversion bias. To provide this background, an analysis based on the depletion approximation that helped to simplify the theories of Schottky barriers and \(p\text{-}n\) junctions will be carried out. The analysis avoids an exact representation of the free-carrier densities in the inversion layer because solutions of the exact case cannot be written in closed form. The complexity of the exact forms is not needed to develop a model of the MOS system that adequately fits our purposes. The complete solutions have been published\(^1\text{,}^3\) and are available for reference.
Thermal-Equilibrium Analysis

For the first level of analysis, it is advantageous to consider the silicon surface region to be in thermal equilibrium with the bulk. Although this assumption will be removed shortly, the results obtained for thermal equilibrium will be enlightening in the more general case. They will also be readily generalized for the nonequilibrium case. We define potential in the silicon as in Equation 4.1.2:

$$\phi(x) = \frac{1}{q} [E_f - E_i(x)]$$  \hspace{1cm} (8.3.1)

Since we consider thermal equilibrium, $E_f$ is constant, but $E_i$ can vary with position. The potential $\phi_p$ in the neutral bulk of the silicon of Figure 8.8 is thus negative because the material is p-type and $E_f$ is less than $E_i$. At the surface, the potential $\phi_s$ is expressed as

$$\phi(0) = \phi_s = \frac{1}{q} [E_f - E_i(0)]$$  \hspace{1cm} (8.3.2)

The carrier densities are related to $\phi(x)$ by Equations 1.1.26 and 1.1.27

$$p = n_i \exp \left( \frac{q \phi}{kT} \right)$$

$$n = n_i \exp \left( \frac{q \phi}{kT} \right)$$  \hspace{1cm} (8.3.3)

From these equations and the definitions of $\phi_p$ and $\phi_s$, we can express the surface free-carrier densities $n_s$ and $p_s$ in terms of the potential drop ($\phi_s - \phi_p$) across the depletion region at the silicon surface.

These equations can be used to obtain the potential at various free-carrier densities in silicon together with corresponding surface-charge conditions at each point.

The strong inversion condition has important applications in device applications. When this condition is reached, the surface potential $\phi_s$ is a sensitive function of $\phi_p$. The values of $\phi_s$ for $N_a$ from 10 to 100 $N_a$ are shown in Figure 8.8. For uniform doping, we can express the surface potential $\phi_s$ in terms of the maximum width of the depletion region at the silicon surface by Equation 3.2.3

$$\phi_s = \frac{q \phi_p}{kT}$$

Equation 8.3.5 can be solved for $\phi_p$.

![Energy-band diagram showing the potential as defined in Equation 8.3.1 in the vicinity of the silicon surface in an MOS system. The sketch corresponds to a positive value of surface potential ($\phi_s$).](image)

Table 8.1 MOS Surface-Charge Conditions

<table>
<thead>
<tr>
<th>$(V_G - V_{FB})$</th>
<th>$\phi_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>Negate $</td>
</tr>
<tr>
<td>0</td>
<td>Negate $</td>
</tr>
<tr>
<td>Positive (small)</td>
<td>Negate $</td>
</tr>
<tr>
<td>Positive (larger)</td>
<td>Positiv $</td>
</tr>
<tr>
<td>Positive (larger)</td>
<td>Positiv $</td>
</tr>
</tbody>
</table>
8.3 MOS Electronics

\[ p_s = N_a \exp \left( \frac{q(\phi_s - \phi_p)}{kT} \right) \]

\[ n_s = \frac{n_i^2}{N_a} \exp \left( \frac{q(\phi_s - \phi_p)}{kT} \right) \]

These equations can be used to give convenient “benchmarks” for the surface potential at various free-carrier densities. Table 8.1 lists these \( \phi_s \) values for \( p \)-type silicon together with corresponding gate-bias ranges and descriptions of the surface-charge conditions at each value of \( \phi_s \).

The strong inversion condition, the last entry in Table 8.1, holds the greatest importance for device applications of the MOS system. Once strong inversion is reached, the surface potential \( \phi_s \) remains relatively constant at \( -\phi_p \) because \( n_s \) is such a sensitive function of \( \phi_s \) (cf. Equation 8.3.4). For example, if \( n_s \) is increased from \( N_a \) to \( 10 \times N_a \), \( \phi_s \) is only increased by 58 mV at room temperature (Problem 8.10). For uniform doping, we can relate the surface potential at the onset of strong inversion to the maximum width of the depletion layer through the use of the depletion approximation as was done for the Schottky barrier in Chapter 3 (Equation 3.2.3)

\[ \phi_s = -\phi_p = \frac{q}{2\varepsilon_s} N_a x_{\text{dmax}}^2 + \phi_p \]  

Equation 8.3.5 can be solved for the maximum width of the depletion layer.

\[ x_{\text{dmax}} = \sqrt{\frac{4\varepsilon_s \phi_p}{qN_a}} \]  

Table 8.1 MOS Surface-Charge Conditions for \( p \)-type Silicon

<table>
<thead>
<tr>
<th>( V_{gs} - V_{fb} )</th>
<th>( \phi_s )</th>
<th>Surface Charge Condition</th>
<th>Surface Carrier Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>Negative</td>
<td>Accumulation</td>
<td>( p_s &gt; N_a )</td>
</tr>
<tr>
<td>(</td>
<td>\phi_s</td>
<td>&gt;</td>
<td>\phi_p</td>
</tr>
<tr>
<td>0</td>
<td>Negative</td>
<td>Neutral (Flat-band)</td>
<td>( p_s = N_a )</td>
</tr>
<tr>
<td>(</td>
<td>\phi_s</td>
<td>=</td>
<td>\phi_p</td>
</tr>
<tr>
<td>Positive (small)</td>
<td>Negative</td>
<td>Depletion</td>
<td>( n_s &lt; p_s &lt; N_a )</td>
</tr>
<tr>
<td>(</td>
<td>\phi_s</td>
<td>&lt;</td>
<td>\phi_p</td>
</tr>
<tr>
<td>Positive (larger)</td>
<td>0</td>
<td>Intrinsic</td>
<td>( p_s = n_s = n_i )</td>
</tr>
<tr>
<td>Positive (larger)</td>
<td>Positive</td>
<td>Weak inversion</td>
<td>( n_i &lt; n_s &lt; N_a )</td>
</tr>
<tr>
<td>(</td>
<td>\phi_i</td>
<td>&lt;</td>
<td>\phi_p</td>
</tr>
<tr>
<td>Positive (larger)</td>
<td>Positive</td>
<td>Onset of strong inversion</td>
<td>( n_s = N_a )</td>
</tr>
<tr>
<td>(</td>
<td>\phi_s</td>
<td>= -\phi_p )</td>
<td></td>
</tr>
<tr>
<td>Positive (larger)</td>
<td>Positive</td>
<td>Strong inversion</td>
<td>( n_s &gt; N_a )</td>
</tr>
<tr>
<td>(</td>
<td>\phi_s</td>
<td>&gt;</td>
<td>\phi_p</td>
</tr>
</tbody>
</table>
The total charge stored in the depletion layer (per unit area) is designated as $Q_d$ where

$$Q_d = -qN_p x_{d_{max}} = -\sqrt{4\epsilon_r qN_p |\phi_p|} \quad (8.3.7)$$

**EXAMPLE  Potential near the Oxide-Silicon Interface**

Derive an expression for the potential distribution in an ideal MOS capacitor in the depletion condition in terms of the surface potential $\phi_s$ and the depletion width at the surface $x_d$ taking the zero for potential in the silicon bulk. The silicon is doped $p$-type and $x = 0$ at the oxide-silicon interface.

**Solution**

Using the depletion approximation and Poisson's equation, the space charge and field gradient are constant and negative in the silicon for $0 \leq x \leq x_d$. Hence, if the surface field is $\mathcal{E}_s$, the field as a function of position away from the surface is

$$\mathcal{E}(x) = \mathcal{E}_s (1 - x/x_d) \quad 0 < x < x_d$$

and

$$\phi(x) = \phi_s \int_0^x \mathcal{E} \; dx$$

$$= \phi_s - \mathcal{E}_s x + \mathcal{E}_s x^2/2x_d$$

$$= \frac{1}{2} \mathcal{E}_s x_d - \mathcal{E}_s x + \mathcal{E}_s x^2/2x_d$$

$$= \frac{\mathcal{E}_s}{2x_d} (x_d - x)^2$$

and

$$\phi(x) = \phi_s \left(1 - \frac{x}{x_d}\right)^2 \quad 0 < x < x_d$$

is the required expression for potential as a function of $x$.

![Diagram showing potential and field distribution](image)

The surface potential $\phi_s$ corresponds to the area under the curve of $\mathcal{E}$ versus $x$. Therefore $\phi_s = \frac{1}{2} \mathcal{E}_s x_d$.

**Nonequilibrium Analysis**

Once the MOS system has reached steady state, the surface and the bulk are at a bias to the $p$-$n$ junction that contacts the inversion layer at the $p$-$n$ interface. The inversion layer at the $p$-$n$ interface is the inversion layer at the $p$-$n$ junction.

An energy-band diagram is characterized by two quasi-linear regions: the $p$-region and one for the $n$-region (Chapter 4). This situation is sketched in the figure below.

An applied reverse bias $V_{bi} - V_n$ is shared between the inversion layer and the $n$-region. The inversion layer for a reverse bias must be applied to the oxide layer when there is a reverse bias at the oxide to silicon interface.
8.3 MOS Electronics

Nonequilibrium Analysis

Once the MOS system has been biased into inversion, a $pn$ junction exists between the surface and the bulk of the silicon. If there is a nearby $n$-type diffused region that contacts the inverted surface as shown in Figure 8.9, it is possible to apply a bias to the $pn$ junction. Application of bias in this manner corresponds to a nonequilibrium condition within the silicon, and some current will flow between the inversion layer at the surface and the bulk. However, in practical applications of MOS systems, the junction will be reverse biased, and the currents will be small.

An energy-band diagram for the case of bias applied to an inverted surface is characterized by two quasi-Fermi levels (Equations 1.1.28 and 1.1.29) one for the $p$-region and one for the $n$-region. Just as for the $pn$ junction under reverse bias (Chapter 4), the two quasi-Fermi levels are separated by the applied reverse bias. This situation is sketched in Figure 8.10 for a reverse bias ($V_C - V_B$) applied between the inversion layer (or channel) and the substrate (or bulk).

An applied reverse bias between the induced surface $n$-region and the bulk increases the charge $Q_d$ in the depletion layer. Since the negative charge induced by $V_C - V_B$ is shared between the depletion and inversion layers, an increase of the charge in the depletion layer means that there is less charge available to form the inversion layer for a given gate voltage. Looked at another way, more gate voltage must be applied to induce the same number of electrons in the inversion layer when there is a reverse bias. With reverse bias present, the surface potential at the onset of strong inversion becomes $\phi_s = -\phi_p + (V_C - V_B)$ rather than

![Figure 8.9](image_url)

Figure 8.9 A diffused junction in the vicinity of an MOS capacitor can be used to bias the induced junction between the bulk of the silicon and an inversion layer formed at the oxide-silicon interface. The cross-hatching indicates the extent of the space-charge region in the depleted silicon.
\( \phi_s = -\phi_p \). The applied voltage extends the range of gate voltages under which the surface region is in depletion (consequently, \( x_{d\text{max}} \) is larger). The bias at the surface prevents an inversion layer from forming as readily as when no bias is present by draining away electrons that could form an inversion layer until the surface potential \( \phi_s \) reaches \( 2|\phi_p| + V_c - V_n \). At this surface potential a channel is formed connected ohmically to the diffused electrode.

In the thermal-equilibrium analysis at the beginning of this section, we noted that \( \phi_s \) does not change appreciably after the surface becomes inverted because the density of free electrons increases exponentially with increasing \( \phi_s \) in inversion. Described in circuit terms, the reverse bias "clamps" the surface potential at a value \( 2|\phi_p| + V_c - V_n \) instead of the unbiased inversion potential \( 2|\phi_p| \). Consequently, the change of the surface potential between flat-band and strong inversion is \( 2|\phi_p| + V_c - V_n \) rather than \( 2|\phi_p| \). The corresponding maximum depletion-region width \( x_{d\text{max}} \) and the depletion-layer charge \( Q_d \) (per unit area) become

\[
x_{d\text{max}} = \sqrt{\frac{2\varepsilon_s(2|\phi_p| + V_c - V_n)}{qN_a}} \tag{8.3.8}
\]

and

\[
Q_d = -\sqrt{2\varepsilon_s q N_a (2|\phi_p| + V_c - V_n)} \tag{8.3.9}
\]

These two quantities are plotted in Figure 8.11 as functions of the dopant concentration for several values of bias between the inversion layer and the bulk.

The previous discussion about charge in the MOS system showed that flat band \( (V_g - V_n = V_{FB}) \) corresponds to the condition of charge neutrality in the silicon. Therefore, \( (V_g - V_n) - V_{FB} \) is the effective voltage tending to charge the MOS capacitor. In this context, the flat-band voltage \( V_{FB} \) for the MOS system is analogous to \( \phi_i \), the built-in voltage for the \( p\!n \) junction; that is, both act like offsets of the zero level in equations relating stored charge to an applied bias. To express the charge in terms of applied voltage, we carry out the following analysis.

The charging voltage \( [V_g - (V_n - V_{FB})] \) and a drop in the silicon \( (V_B) \) are

\[
\varepsilon_{ax} = V_{ax} / \varepsilon_{ax0} \tag{8.3.10}
\]

The field in the insulating oxide, in terms of the applied voltage \( V_{ax} \) is

\[
\varepsilon_{ax} = V_{ax} / \varepsilon_{ax0} \tag{8.3.10}
\]

Just inside the silicon and oxide (as encountered), the normal capacitance \( C_{ox} \) is

\[
C_{ox} = \varepsilon_{ox} / \varepsilon_{ox0} \tag{8.3.11}
\]

If Equation 8.3.12 is compared with the following equation (per unit area)

\[
\varepsilon_{ax} \phi_{ax0} = \varepsilon_{ox} \phi_{ox0} \tag{8.3.13}
\]

Gauss' law states that the electric field is constant (times the electric field in a field-free bulk region, we...
The charging voltage \( [V_G - V_b] - V_{FB} \) is the sum of a drop \( V_{ox} \) across the oxide and a drop in the silicon (\( \phi_s - \phi_p \)) (cf. Figure 8.6d).

\[
V_G - V_b - V_{FB} = V_{ox} + \phi_s - \phi_p \tag{8.3.10}
\]

The field in the insulating oxide is constant in the absence of any oxide charge. In terms of the applied voltage and oxide thickness, this field \( \delta_{ox} \) is

\[
\delta_{ox} = \frac{V_{ox}}{x_{ox}} = \frac{[V_G - V_b - V_{FB} - (\phi_s - \phi_p)]}{x_{ox}} \tag{8.3.11}
\]

Just inside the silicon and adjacent to the oxide (before any silicon charge is encountered), the normal displacement \( D \) will be constant and the field \( \delta_{so} \) will be

\[
\delta_{so} = \frac{\varepsilon_{ox}\delta_{ox}}{\varepsilon_i} \tag{8.3.12}
\]

If Equation 8.3.12 is combined with Equation 8.3.11 and the definition for oxide capacitance (per unit area) \( C_{ox} = \varepsilon_{ox}/x_{ox} \) (Equation 8.2.1) is used, we find

\[
\varepsilon_i\delta_{so} = C_{ox}[V_G - V_b - V_{FB} - (\phi_s - \phi_p)] \tag{8.3.13}
\]

Gauss' law states that the charge contained in a volume equals the permittivity times the electric field emanating from the volume. Applying Gauss' law to a volume extending from just inside the silicon at the oxide-silicon interface to the field-free bulk region, we can write

\[-\varepsilon_i\delta_{so} = Q_s = Q_n + Q_d \tag{8.3.14}\]
where $Q_s$, the total charge induced in the semiconductor, is composed of the mobile electron charge $Q_n$ and the depletion-region charge $Q_d$ (in all per unit area). These quantities are shown in Figure 8.6b. Using Equation 8.3.14 in Equation 8.3.13, we can express the mobile charge $Q_n$ as

$$Q_n = -C_{ox}(V_G - V_{FB} - V_B) - (\phi_s - \phi_p) - Q_d$$  (8.3.15)

We insert the values of $\phi_s$ and $Q_d$ into Equation 8.3.15 in order to relate the mobile charge $Q_n$ to the applied voltages. In strong inversion with an applied reverse bias between the channel and the bulk, $\phi_s = -\phi_p + (V_C - V_B)$, and Equation 8.3.15 becomes

$$Q_n = -C_{ox}(V_G - V_{FB} - V_B - 2|\phi_p|) + \sqrt{2\varepsilon\varepsilon_0 N_d}(2|\phi_p| + V_C - V_B)$$  (8.3.16)

Equation 8.3.16 reduces to the simpler expression

$$Q_n = -C_{ox}(V_G - V_{FB} - V_B - 2|\phi_p|) + \sqrt{4\varepsilon\varepsilon_0 N_d|\phi_p|}$$  (8.3.17)

with no applied bias between the channel and the bulk. Note that the first terms on the right-hand side of either Equation 8.3.16 or 8.3.17 are negative while the second terms in both equations are positive. The positive terms are, however, smaller in magnitude (since $|Q_s| > |Q_d|$), so that the difference $Q_n$ is negative, as must be the case for the inversion layer on a p-type substrate.

From these equations, we can directly express the gate voltage necessary to induce a conducting channel at the surface of the semiconductor. This voltage, known as the threshold voltage $V_T$, is defined as the gate voltage that results in $Q_n = 0$. From Equation 8.3.16, an expression for $V_T$ can be written as

$$V_T = V_{FB} + V_C + 2|\phi_p| + \frac{1}{C_{ox}} \sqrt{2\varepsilon\varepsilon_0 N_d(2|\phi_p| + V_C - V_B)}$$  (8.3.18)

Taken one by one, the terms in Equation 8.3.18 are readily seen to have the proper qualitative behavior. First, $V_T$ contains $V_{FB}$ because a gate voltage equal to $V_{FB}$ is necessary to bring the silicon to a charge-neutral condition. (In the system of Figure 8.2, $V_{FB}$ is negative, tending to reduce the size of $V_T$.) Second, increasing the channel voltage $V_C$ increases the gate voltage necessary to induce a given charge near the silicon surface. Third, $2|\phi_p|$ volts must be applied to cause the silicon energy bands to be bent to an inverted condition. Finally, the square-root term accounts for the uniform distribution of space charge in the depletion region. This term is inversely proportional to the oxide capacitance. It increases with $V_C - V_B$ to reflect the redistribution of semiconductor charge $Q_n$ from the inversion layer (where it contributes to $Q_n$) into the depletion layer (where it forms part of $Q_d$).

The charge in the inversion layer can be expressed very simply in terms of the difference between the applied gate voltage and the threshold voltage. From Equations 8.3.16 and 8.3.18

$$Q_n = -C_{ox}(V_G - V_T)$$  (8.3.19)

Equation 8.3.19 should be used with caution for $V_G$ approximating equal to $V_T$, because it is derived on the assumption that there are no electrons at the surface until the onset of strong inversion. It can be safely applied to large values of $V_G$ and $V_T$. This is, of course, an approximation since solutions of Poisson's equation have been collected for MOS transistors in the special case of strong inversion. The more frequent case of strong inversion and weak inversion may therefore have been collected for general reference. Results are included.

The theory presented here is characteristic of the oxide-semiconductor interface and the silicon interface. At its interface, the oxide-semiconductor interface is unimportant, and the oxide-semiconductor interface is unimportant. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important. The oxide-semiconductor interface is thus important.
8.4 Oxide and Interface Charge

until the onset of strong inversion when the surface potential \( \phi_s = -\phi_p + V_C - V_B \).

This is, of course, an approximation that can be removed by considering the exact solutions of Poisson's equation for the MOS system.¹,³

The more frequently used equations that we have derived for the MOS system have been collected in Table 8.2 which is included at the end of the chapter for ready reference. Results for n-type substrates as well as p-type substrates are included.

8.4 Oxide and Interface Charge

The theory presented thus far has neglected consideration of an important characteristic of the oxide-silicon system. This is the influence of charge within the oxide and at its surfaces. The presence of charge in the oxide and at the oxide-silicon interface is unavoidable in practical systems. To appreciate the importance of oxide charge, it is worthwhile to estimate the order of magnitude of the charge densities that have been considered in discussing the MOS system. Just above the threshold voltage where the system enters the inversion region, for example, the surface density of electrons \( Q_d/q \) will be of the same order of magnitude as the density of dopant atoms (per unit area). For homogeneously distributed dopant atoms, the area density is \( N_a \), or \( 10^{10} \text{ cm}^{-2} \) if \( N_a \) is \( 10^{15} \text{ cm}^{-3} \). When this density is compared to the area density of silicon atoms (\( 5 \times 10^{22} \) or \( 1.35 \times 10^{15} \text{ cm}^{-2} \)), it becomes clear that a surface-charge density only about \( 10^{-5} \) times as large as the atom density can cause the MOS system to depart from the ideal analysis. Fortunately, when they are formed carefully, interfaces between thermally grown, amorphous silicon dioxide and single-crystal silicon can contain charge densities of the order of \( 10^{10} \text{ cm}^{-2} \) or lower. This low charge density is unique and is one of the outstanding attributes of the oxide-silicon system. We shall consider specific sources for oxide charge after we have calculated its influence on MOS properties.

Theoretical Analysis. Consider that a density of charge \( Q_{ox} \) is located at the plane \( x = x_1 \) within the oxide as shown in Figure 8.12a. The charges at \( x_1 \) will induce equal and opposite charges that are divided, in general, between the silicon and the metal gate. The closer is \( x_1 \) to \( x_{ox} \), the oxide-silicon interface, the greater will be the fraction of induced charge within the silicon. Because this induced charge changes the charge stored in the silicon at thermal equilibrium, it correspondingly alters the flat-band voltage from the value predicted in the ideal MOS analysis (Equation 8.1.1).

The size of the shift in the flat-band voltage is readily found by using Gauss' law to obtain the value of gate voltage that causes all of the oxide charge \( Q_{ox} \) to be mirrored in the gate electrode—so that none is induced in the silicon. This condition is illustrated in Figure 8.12.b. Referring to the figure and employing Gauss' law, we note that the field is constant between the metal (at \( x = 0 \)) and \( Q_{ox} \) (at \( x_1 \)) and zero between \( x_1 \) and the silicon surface (at \( x = x_{ox} \)). Its value \( \delta_{ox} \)
An expression for the gate voltage resulting from the presence of $Q_{ox}$ is the negative integral of $\delta_{ox}$. Because it contributes to the flat-band voltage, we call it $\Delta V_{FB}$ (i.e., the change in $V_{FB}$ (Equation 8.1.1) from the ideal MOS case because of an interfacial oxide charge). The field and potential variations for the zero-bias case (Figure 8.12a) are shown in Figure 8.12c; those for flat band (Figure 8.12b) are shown in Figure 8.12d.

\[
\delta_{ox} = -\frac{Q_{ax}}{e_{ox}} \quad 0 < x < x_1 \quad (8.4.1)
\]

The maximum value of $C_{ox}$ the oxide capacitance at the silicon interface ($x_1$) is accounted for in the shift in flat-band voltage throughout the oxide.$\Delta V_{FB}$

Fixed charge at the interface that is incorporated into the above expression for gate and in the silicon by combining Equation 8.1.1 with the shift in flat-band voltage from flat-band procedure (Equation 8.3.18). Expressions for gate and in the silicon by combining Equation 8.1.1 with the shift in flat-band voltage from flat-band procedure (Equation 8.3.18).

In some cases, oxide that can be influenced by the presence of interface charge as well as the interfacial oxide charge.

- The fixed interface-cha
An expression for $\Delta V_{FB}'$ is

$$\Delta V_{FB}' = x_1 \varepsilon_{ox} = -\frac{x_1 Q_{ox}}{C_{ox}}$$  \hspace{1cm} (8.4.2)

Equation 8.4.2 can be rewritten by using Equation 8.2.1 to express $\Delta V_{FB}'$ in terms of $C_{ox}$, the oxide capacitance per unit area.

$$\Delta V_{FB}' = -\frac{Q_{ox} x_1}{C_{ox} \varepsilon_{ox}}$$  \hspace{1cm} (8.4.3)

The maximum value for $\Delta V_{FB}'$ occurs expectedly when $Q_{ox}$ is situated at the oxide-silicon interface ($x_1 = x_{ox}$), because the charge induced by $Q_{ox}$ is then contained entirely in the silicon. In contrast, when $Q_{ox}$ is adjacent to the gate, it has no effect on $\Delta V_{FB}'$.

The results given for the sheet of charge at $x = x_1$ can be generalized to account for the shift in flat-band voltage by an arbitrary distribution of charge $\rho(x)$ by superposing and integrating the increments that result from charges distributed throughout the oxide. The overall result is

$$\Delta V_{FB} = \frac{1}{C_{ox}} \int_0^{x_{ox}} \frac{x}{x_{ox}} \rho(x) dx$$  \hspace{1cm} (8.4.4)

Fixed charge at the oxide-silicon interface is frequently treated separately from charge that is incorporated in the oxide itself, even though surface charge can be accounted for in the formulation of Equation 8.4.4. The fixed interface charge density is designated $Q_f^*$, and its contribution to the flat-band voltage is

$$\Delta V_{FB}' = -\frac{Q_f^*}{C_{ox}}$$  \hspace{1cm} (8.4.5)

An expression for $V_{FB}$ that includes the effects of differing work functions in the gate and in the silicon, as well as the influence of fixed oxide charge, can be written by combining Equations 8.1.1, 8.4.4 and 8.4.5.

$$V_{FB} = \Phi_{MS} - \frac{Q_f}{C_{ox}} - \frac{1}{C_{ox}} \int_0^{x_{ox}} \frac{x}{x_{ox}} \rho(x) dx$$  \hspace{1cm} (8.4.6)

From Equation 8.4.6, we see that the effect of oxide charge is to shift the flat-band voltage from its value in the ideal case. If the oxide charge is stable, the shift in flat band produces a corresponding shift in the threshold voltage $V_T$ (Equation 8.3.18). Experimentally, this threshold shift would cause the capacitance versus gate-voltage curves to be translated along the $V_T$ axis. A typical result for a high-frequency capacitance curve is sketched in Figure 8.13 (dashed curve).

In some cases, oxides and oxide-silicon interfaces can contain unstable charges that can be influenced by the applied voltage. In this case, the threshold voltage is itself dependent on the gate voltage. The capacitance versus voltage curve will then be distorted as in the dotted curve in Figure 8.13. To understand this behavior as well as the influence of fixed charge, we discuss the physical sources of oxide charge.

* The fixed interface-charge density is also often called $Q_{ox}$. 

---

Note: The text continues with further explanations and derivations related to oxide and interface charge.
MOSFET

8

8.1 INTRODUCTION

The metal–oxide–semiconductor field-effect transistor (MOSFET) is the most important device for very-large-scale integrated circuits such as microprocessors and semiconductor memories. MOSFET is also becoming an important power device. It has many acronyms including IGFET (insulated-gate field-effect transistor) MISFET (metal-insulator-semiconductor field-effect transistor) and MOST (metal–oxide–semiconductor transistor). The principle of the surface field-effect transistor was first proposed in the early 1930s by Lilienfeld and Heil. It was subsequently studied by Shockley and Pearson in the late 1940s. In 1960, Kahng and Atalla proposed and fabricated the first MOSFET using a thermally oxidized silicon structure. The basic device characteristics have been subsequently studied by Itholola and Moll, Sah, and Hofstein and Heiman. The technology, application, and device physics have been reviewed by Wallmark and Johnson, Richman, and Brews.

Because the current in a MOSFET is transported predominantly by carriers of one polarity only (e.g., electrons in an n-channel device), the MOSFET is usually referred to as a unipolar device. The MOSFET is a member of the family of field-effect transistors. The other members, JFETs and MESFETs, have already been considered in Chapter 6. Al-
though MOSFETs have been made with various semiconductors such as Ge, Si, and GaAs, and use various insulators such as SiO₂, Si₃N₄, and Al₂O₃, the most important system is the Si-SiO₂ combination. Hence most of the results in this chapter are obtained from the Si-SiO₂ system.

We first consider the basic device characteristics of the so-called long-channel MOSFET; that is, the channel length \( L \) is much longer than the sum of the source and drain depletion-layer widths \( (W_S + W_D) \). This serves as a foundation to understand short-channel, that is, \( L \leq (W_S + W_D) \), and related MOSFET devices.

Figure 1 shows the reduction of the minimum device dimension since the beginning of the integrated circuit era in 1959. Figure 1 also shows that the minimum dimension will shrink continuously; the 1-μm barrier for commercial devices may be overcome by 1990. The reduction of device dimensions is driven by the requirement that integrated circuits of high complexity be fabricated. The number of components per integrated-circuit chip has grown exponentially since 1959 (Fig. 2). The rate of growth is expected to slow down because of a lack of product definition and design. However, a complexity of 1 million or more devices per chip may be available around 1990 using 1-μm or submicron device geometries. As the channel length becomes shorter, one has to consider short-channel effects due to two-dimensional potential, high-field transport and oxide charging. Many device structures have been proposed to improve MOSFET performance. Some representative structures as well as the nonvolatile semiconductor memory, basically a MOSFET with a multilayer gate structure, will be discussed.

8.2 BASIC DEVICE CHARACTERISTICS

The basic structure of a metal-oxide-semiconductor field-effect transistor (MOSFET) is illustrated in Fig. 3. It is a four-terminal device and consists of a \( p \)-type semiconductor substrate into which two \( n^+ \) regions, the source and drain, are formed (e.g., by ion implantation). The metal contact on the insulator is called gate; heavily doped polysilicon or a combination of silicide and polysilicon can also be used as the gate electrode. The basic device parameters are the channel length \( L \), which is the distance between the two metallurgical \( n^+ \)-\( p \) junctions; the channel width \( W \); the insulator thickness \( d \); the junction depth \( r \); and the substrate doping \( N_A \). In a silicon integrated circuit, a MOSFET is surrounded by a thick oxide (called the field oxide to distinguish it from the gate oxide) to isolate it from adjacent devices.

The source contact will be used as the voltage reference throughout this

\*These terms will be defined in Section 8.2.

\*This is an \( n \)-channel device; one may consider a \( p \)-channel device by exchanging \( p \) for \( n \) and reversing the polarity of the voltage.
chapter. When no voltage is applied to the gate, the source-to-drain electrodes correspond to two p-n junctions connected back to back. The only current that can flow from source to drain is the reverse leakage current.\(^*\) When a sufficiently large positive bias is applied to the gate so that a surface inversion layer (or channel) is formed between the two n\(^+\) regions, the source and the drain are then connected by a conducting-surface n channel through which a large current can flow. The conductance of this channel can be modulated by varying the gate voltage. The back-surface contact (or substrate contact) can have the reference voltage or be reverse-biased; the back-surface voltage will also affect the channel conductance.

8.2.1 Nonequilibrium Condition

When a voltage is applied across the source–drain contacts, the MOS structure is in a nonequilibrium condition; that is, the imref of the minority carriers (electrons, in the present case) is lowered from the equilibrium Fermi level. To show more clearly the band bending across the device, Fig. 4a shows the MOSFET turned 90\(^\circ\). The two-dimensional, flat-band, zero-bias \((V_G = V_D = V_{BS} = 0)\) equilibrium condition is shown in Fig. 4b. The equilibrium conditions under a gate bias that causes surface inversion are shown in Fig. 4c. The nonequilibrium condition with both drain and gate biases is shown in Fig. 4d, where we note the separation of the imref of electrons and holes; the hole imref \(E_{Fp}\) remains at the bulk Fermi level while the electron imref \(E_{Fn}\) (minority in the present case) is lowered toward the drain contact. Figure 4d shows that the gate voltage required for inversion at the drain is larger than the equilibrium case in which \(\psi_{(inv)} = 2\psi_b\). This is because the applied drain bias lowers the electron imref. and an inversion layer can be formed only when the potential at the surface crosses over the imref of the minority carrier.

Figure 5 shows a comparison\(^*\) of the charge distribution and energy-band variation of an inverted p region for the equilibrium case and the nonequilibrium case at the drain. For the equilibrium case (discussed in Chapter 7), the surface depletion region reaches a maximum width \(W_s\) at
inversion. For the nonequilibrium case, the depletion-layer width is a function of the bias $V_D$, and the surface potential $\psi$, at the onset of strong inversion is given, to a good approximation, by

$$\psi_{\text{(inv)}} = V_D + 2\psi_B.$$  

(1)

The derivation for the characteristic of the surface-space-charge under the nonequilibrium condition is similar to that in Chapter 7. The two assumptions are that (1) the irref for the majority carriers of the substrate does not vary with distance from the bulk to the surface, and (2) the irref for the minority carriers of the substrate is separated by the applied junction bias $V_D$ from the irref for the majority carriers; that is, $E_{Fm} = E_{Fp} + qV_D$ for a $p$ substrate. The first assumption introduces little error when the surface is inverted, because majority carriers are then only a negligible part of the surface space charge; the second assumption is correct under the inversion condition, because minority carriers are an important part of the surface-space-charge region when the surface is inverted. Based on these assumptions, the one-dimensional Poisson equation for the surface-space-charge region at the drain is given by

$$\frac{\partial^2 \psi}{\partial x^2} = -\frac{q}{\varepsilon_s} (N_D - N_A + p - n)$$

(2)

where

$$N_D - N_A = n_{po} - p_{po}, \quad p_{po} = N_A$$

$$p = p_{po} e^{-\beta\psi}, \quad n = n_{po} e^{\beta\psi - \psi_0}, \quad \beta = q/kT.$$  

(3)

Following the same approach as in Chapter 7, we obtain

$$\xi = \frac{-\partial \psi}{\partial x} = \sqrt{\frac{kT}{qL_D}} F(\beta\psi, V_D, \frac{n_{po}}{p_{po}})$$

(4)

and

$$Q_s = -\xi \xi' = \sqrt{\frac{2e_kT}{qL_D}} F(\beta\psi, V_D, \frac{n_{po}}{p_{po}})$$

(5)

where

$$F(\beta\psi, V_D, \frac{n_{po}}{p_{po}}) = \left[e^{-\beta\psi} + \beta\psi - 1 + \frac{n_{po}}{p_{po}} e^{-\psi_0} (e^{\beta\psi} - \beta\psi e^{\beta\psi_0} - 1)\right]^{1/2}$$

(6)

and

$$L_D = \left(\frac{kT}{e_kq}\right)^{1/2}.$$  

(7)

The surface charge per unit area after strong inversion is given by

$$Q_s = Q_a + Q_B$$

(8)

where

$$Q_B = -qN_A W_m = -\sqrt{2qN_A \varepsilon_s (V_D + 2\psi_B)}$$

(9)

and $Q_a$, the charge due to minority carriers within the inversion layer, is

$$|Q_a| = \int_{x_0}^{x_s} n(x) dx = \int_{x_0}^{x_s} n(\psi) d\psi$$

(10)

or

$$|Q_a| = \int_{x_0}^{x_s} \left(\frac{1}{\sqrt{2kT/qL_D}} F(\beta\psi, V_D, \frac{n_{po}}{p_{po}})\right) d\psi$$

(11)

where $x_s$ denotes the point at which the intrinsic Fermi level intersects the irref for electrons. For the practical doping ranges in silicon, the value of $x_s$ is quite small, of the order of 30 to 300 Å. Equation 11 is the basic formula for long-channel MOSFET, and can be evaluated numerically.
Under strong inversion conditions, a simplified expression for \( Q_s \) can be obtained from a charge-sheet model\(^9\) and is given by

\[
|Q_s| = \sqrt{2qN_sL_0 \left[ \beta_0 \psi_0 + \left( \frac{p_{eq}}{p_{pn}} \right) e^{\text{sat}}. u_{eq} \right] }^{1/2} - (\beta_0)^{1/2}.
\]  

(12)

This expression for \( Q_s \) is derived under the condition \( V_{FE} = 0 \). When a substrate reverse bias is applied, the depletion width increase, and the term \( \beta V_D \) in Eq. 12 is replaced by \( \beta (V_D + V_{BS}) \).

### 8.2.2 Linear and Saturation Regions

We shall first present a qualitative discussion of device operation. Let us consider that a voltage is applied to the gate, causing an inversion at the semiconductor surface, Fig. 6a. If a small drain voltage is applied, a current will flow from the source to the drain through the conducting channel. Thus the channel acts as a resistance, and the drain current \( I_D \) is proportional to the drain voltage \( V_D \). This is the linear region. As the drain voltage increases, \( I_D \) eventually reaches a point at which the channel depth \( x \), at \( y = L \), is reduced to zero; this is called the pinch-off point, Fig. 6b. Beyond the pinch-off point the drain current remains essentially the same, because for \( V_D > V_{D,sat} \), the voltage at \( Y \) remains the same, \( V_{D,sat} \). Thus the number of carriers arriving at point \( Y \) from the source, and hence the current flowing from source to drain, remains the same apart from a decrease in \( L \) to the value \( L' \) (Fig. 6c). Carrier injection from \( Y \) into the drain-depletion region is quite similar to the case of carrier injection from an emitter–base junction to the base–collector depletion region of a bipolar transistor.

We shall now derive the basic MOSFET characteristics under the following idealized conditions: (1) the gate structure corresponds to an ideal MOS diode as defined in Chapter 7; that is, there are no interface traps, fixed oxide charge, or work-function difference, and so on; (2) only drift current will be considered; (3) carrier mobility in the inversion layer is constant; (4) doping in the channel is uniform; (5) reverse leakage current is negligibly small; and (6) the transverse field \( (\vec{E}_x \text{ in the } x \text{ direction}) \) in the channel is much larger than the longitudinal field \( (\vec{E}_y \text{ in the } y \text{ direction}) \). The last condition corresponds to the so-called gradual channel approximation.

Under such idealized conditions, the total charge induced in the semiconductor per unit area \( Q \), at a distance \( y \) from the source is given by

\[
Q_s(y) = \frac{1}{C} \left[ -V_s + \psi_s(y) \right] C.
\]

(13)

where \( C = \varepsilon / d \) is the capacitance per unit area. The charge in the inversion layer is given by

\[
Q_i(y) = Q_s(y) - Q_n(y) = \frac{1}{C} \left[ -V_s + \psi_s(y) \right] C.
\]

(14)

The surface potential \( \psi_s(y) \) at inversion can be approximated by \( 2\psi_0 + \)

\[ V(y), \text{ where } V(y) \text{ is the reverse bias between point } y \text{ and the source electrode (which is assumed to be grounded). The charge within the surface depletion region } Q_d(y) \text{ was given previously as } \]

\[
Q_d(y) = -qN_sW_m = -\sqrt{2\varepsilon_0 qN_s [V(y) + 2\psi_0]}.
\]

(15)

Substituting Eq. 15 into Eq. 14 yields

\[
Q_s(y) = \frac{1}{C} \left[ -V_s - V(y) + 2\psi_0 \right].
\]

(16)
The conductivity of the channel can be approximated by
\[
\sigma(x) = qn(x)\mu_s(x).
\] (17)

The channel conductance is then given by
\[
g = \frac{Z}{L} \int_0^1 \sigma(x) \, dx.
\] (18)

For a constant mobility, the channel conductance becomes
\[
g = \frac{qZ\mu_s}{L} \int_0^1 n(x) \, dx = qZ\mu_s|Q_s/L|.
\] (19)

The channel resistance of an elemental section \(dy\), Fig. 6a, is given by
\[
dR = \frac{dy}{gL} = \frac{dy}{Z\mu_s|Q_s(y)|}
\] (20)

and the voltage drop across this elemental section is given by
\[
dV = I_d \, dr = \frac{I_d \, dy}{Z\mu_s|Q_s(y)|}
\] (21)

where \(I_d\) is the drain current and is a constant independent of \(y\). Substituting Eq. 16 into Eq. 21 and integrating from the source \((y = 0, \ V = 0)\) to the drain \((y = L, \ V = V_D)\) yields
\[
I_d = \frac{Z}{L} \mu_s C_i \left\{ \left( V_D + 2\psi - \frac{V_D}{2} \right) V_D - \frac{2\sqrt{2}\varepsilon \eta n_s}{3C_i} \left( (V_D + 2\psi)^{1/2} - (2\psi)^{1/2} \right) \right\}
\] (22)

for the present idealized case.

Equation 22 predicts that for a given \(V_e\), the drain current first increases linearly with drain voltage (the linear region), then gradually levels off, approaching a saturated value (the saturation region). The basic output characteristic of an idealized MOSFET is shown in Fig. 7. The dashed line indicates the locus of the drain voltage \((V_D)\) at which the current reaches a maximum value.

We shall now consider the above-mentioned two regions. For the case of small \(V_D\), Eq. 22 reduces to
\[
I_d = \frac{Z}{L} \mu_s C_i \left\{ \left( V_D - V_t \right) V_D - \left( \frac{\varepsilon \eta n_s \sqrt{V_D}}{4C_i} \right) V_D^2 \right\}
\] (23)

or
\[
I_d = \left( \frac{Z}{L} \right) \mu_s C_i (V_D - V_t) V_D \quad \text{for} \quad V_D \ll (V_D - V_t)
\] (23a)

where \(V_t\) (the threshold voltage) is given by
\[
V_t = 2\psi - \frac{\sqrt{2\varepsilon \eta n_s (2\psi)}}{C_i}.
\] (24)

Fig. 7 Idealized drain characteristics \((I_d \ vs V_D)\) of a MOSFET. The dashed line indicates the locus of the saturation drain voltage \((V_D)\) at which \(V_D > V_{Dsat}\), the drain current remains constant.

The calculated values of \(V_I\) as a function of semiconductor doping density and insulator thickness were shown in Chapter 7 for the Si-SiO2 system. By plotting \(I_d \ vs \ V_D\) (for a given small \(V_D\)), the threshold voltage can be deduced from the linearly extrapolated value at the \(V_D\) axis. In the linear region, Eq. 23a, the channel conductance \(g_d\) and the transconduction \(g_m\) are given as
\[
g_d = \frac{\partial I_d}{\partial V_D} \bigg|_{V_D \to 0} = \frac{Z}{L} \mu_s C_i (V_D - V_t)
\] (25)

\[
g_m = \frac{\partial I_d}{\partial V_D} \bigg|_{V_D \to V_D} = \frac{Z}{I} \mu_s C_i V_D
\] (26)

When the drain voltage is increased to a point such that the charge in the
inversion layer \( Q(y) \) at \( y = L \) becomes zero, the number of mobile electrons at the drain experiences a drastic fall-off. This point, called pinch-off, is analogous to the junction field-effect transistor. The drain voltage and the drain current at this point are designated as \( V_{D,\text{sat}} \) and \( I_{D,\text{sat}} \), respectively.

Beyond the pinch-off point we have the saturation region. The value of \( V_{D,\text{sat}} \) is obtained from Eq. 16 under the condition \( Q_s(L) = 0 \):

\[
V_{D,\text{sat}} = V_G - 2\psi_B + K \left( 1 - \sqrt{1 + 2V_d/K} \right)
\]

(27)

where \( K = \sqrt{\varepsilon_i q N_s / C_i} \). The saturation current \( I_{D,\text{sat}} \) can be obtained by substituting Eq. 27 into Eq. 22:

\[
I_{D,\text{sat}} = \frac{mZ}{L} \mu_a C_i (V_G - V_T)^3.
\]

(28)

where \( m \) is a function of doping concentration and approaches \( \frac{1}{3} \) at low dopings.\(^{11}\)

The threshold voltage \( V_T \) in the saturation region is the same as given by Eq. 25 for low substrate dopings and thin insulator layers. For higher dopings, \( V_T \) becomes \( V_G \)-dependent. The transconductance in the saturation region when Eq. 28 applies is given by

\[
g_m = \left. \frac{\partial I_D}{\partial V_G} \right|_{V_D = \text{const}} = \frac{2mZ}{L} \mu_a C_i (V_G - V_T).
\]

(29)

In previous discussions, we made many assumptions to bring out the most important characteristics of the MOSFET. We shall now remove the first two assumptions and consider the effects due to a nonideal gate MOS and diffusion current. The main effect of the fixed oxide charges and the difference in work functions is to cause a voltage shift corresponding to the flat-band voltage \( V_{FB} \). This in turn causes a change in the threshold voltage \( V_T \); in the linear region \( V_T \) becomes

\[
V_T = V_{FB} + 2\psi_B + \frac{\sqrt{2\varepsilon_i q N_s (2\psi_B)}}{C_i} \\
= \left( \phi_m - \frac{Q}{C_i} \right) + 2\psi_B + \frac{\sqrt{4\varepsilon_i q N_s \psi_B}}{C_i}
\]

(30)

When a substrate bias is applied, the threshold voltage becomes

\[
V_T = V_{FB} + 2\psi_B + \sqrt{2\varepsilon_i q N_s (2\psi_B + V_{BS})}/C_i,
\]

(31)

or

\[
\Delta V_T = V_T(V_{BS}) - V_T(V_{BS} = 0) \\
= \frac{\sqrt{2\varepsilon_i q N_s}}{C_i} \left( 2\psi_B + 2V_{BS} - \sqrt{2\psi_B} \right) \\
= \frac{a}{\beta} \left( \sqrt{2\beta \psi_B} + \beta V_{BS} - \sqrt{2 \psi_B} \right)
\]

(32)

In Fig. 8, oxide thickness versus substrate doping is plotted for given \( a \) values\(^{19}\) using Eq. 33. The \( a \) values increase with increasing doping and oxide thickness.

\[
a = \sqrt{2(\varepsilon_i/L_D)/C_i} = 2(\varepsilon_i/\varepsilon_0)(d/L_D).
\]

(33)

Threshold voltage shift versus \( V_{BS} \) is plotted in Fig. 9 for various \( a \) values. As the \( a \) value increases, \( \Delta V_T \) also increases. For a given \( a \) value, the resulting variation in \( \Delta V_T \) is indicated by vertical bars for substrate dopings ranging from \( 10^{19} \) to \( 10^{17} \) cm\(^{-2} \) (Fig. 9). The primary influence upon \( \Delta V_T \) is the choice of \( a \) itself; the influence of doping or oxide thickness upon \( \Delta V_T \), independent of \( a \), is minor.

To consider the effect of the diffusion current component, we refer to Fig. 4 for the nonequilibrium condition. The drain current density including
total drain current based on the gradual-channel approximation is

$$I_D = \int_0^L J_D(x, y) Z \, dx$$

$$= \frac{1}{L} \int_0^L D_y q Z \left( \frac{\partial \psi_m}{\partial y} \right) \int_0^L n(x, y) \, dx \, dy$$

$$= \frac{Z \varepsilon_s \mu_n}{L} \int_0^{V_D} \int_0^{\psi_m} \frac{e^{d\psi - \psi V}}{F(\psi_m, V, n_{eq}, p_{eq})} \, d\psi \, dV.$$  \hspace{1cm} (35)

The gate voltage $V_G$ is related to the surface potential $\psi_s$ by

$$V_G = V_D - V_{FB} = - \frac{Q}{C_i} + \psi_s$$

$$= \frac{2\varepsilon_s kT}{C_i q L_D} F(\beta \psi_s, V, n_{eq}, p_{eq}) + \psi_s. \hspace{1cm} (36)$$

Equation 35 reduces to Eq. 22 for gate voltages well above threshold. Equation 22 however, becomes inaccurate for gate voltages near threshold and near pinch-off. For a particular device with known physical dimensions, bulk impurity concentration, and effective mobility, Eq. 35 can be calculated numerically to give accurate results for the entire range of drain voltage from the linear region to the saturation region. Figure 10 demonstrates the current saturation phenomena very well, showing a typical drain characteristic for a long-channel MOSFET.

Fig. 9 Threshold voltage shift versus substrate reverse bias for various a values.

both drift and diffusion components is given by

$$J_D(x, y) = q\mu_n n(x, y) + qD_z \nabla n$$

$$= -qD_z n(x, y) \nabla \psi_m. \hspace{1cm} (34)$$

where $\psi_m$ is the electron imref measured from the bulk Fermi level. The

Fig. 10 Theoretical (dots) and experimental (solid lines) drain characteristics of a p-channel MOSFET having $d = 2000$ A, $\mu_n = 4 \times 10^4$ cm$^2$/Vs, and $\mu_e = 256$ cm$^2$/Vs (After Pino and Sah, Ref. 16).
channel. In the JFET, this channel is defined by $pn$-junction depletion regions that vary in extent when the reverse bias on the junction changes. In the MOSFET, the density of free charge in the channel is controlled by a field that extends from the gate electrode to the silicon through the insulating layer. The properties of the MOS system that were described in Chapter 8 are the framework for the electronics of the MOSFET.

The symbols that have been adopted for MOSFETs are sketched in Figure 9.2a and b for $p$-channel enhancement and depletion devices, and in Figure 9.2c and d for $n$-channel enhancement and depletion devices. Note that $n$- and $p$-channel MOSFETs are differentiated by an arrow that represents the $pn$ junction between the substrate and induced channel at the surface. In many circuit diagrams the connection to the substrate is not indicated on the symbol because it is evident to the designer. The symbol for depletion-mode MOSFETs has a heavy line between the source and drain, indicating the presence of a channel when no voltage is applied between the source and the gate. Unless an asymmetric geometry is used, the MOSFET is a bilateral device at its output terminals, and the source and drain electrodes are drawn identically.

### 9.1 Basic Theory

The $n$-type source and drain diffusions in the MOSFET shown in Figure 9.1 are separated by a lateral distance known as the channel length ($L$). The channel length extends along the $y$-axis in the figure, while the direction into the silicon (perpendicular to the oxide) is conventionally the $z$-direction. We shall see that there are advantages to making $L$ small; in typical MOSFETs, channel lengths are of the order of a few micrometers, with minimum lengths determined mainly by limitations in lithography. The channel width ($W$) is perpendicular to the figure (in the $z$-direction). Channel widths are selected to meet circuit-design requirements. The circuit designer specifies a value for $W$ to achieve a desired conductance for a given bias on the MOSFET. The thickness of the oxide ($\tau_{ox}$) separating the gate from the channel is typically about 40 nm although it can range from roughly one fourth to four times this value.
The MOSFET is a four-terminal device with connections to the source, drain, gate, silicon substrate (or bulk). The voltages at these terminals are shown in Figure 9.1 with appropriate subscripts. In this chapter, the MOSFET analysis is carried out without assigning any of the terminal voltages to be a reference. All voltages are therefore expressed in the equations we derive to retain maximum flexibility.

### 9.1 Basic Theory

The MOSFET shown in Figure 9.1 are depletion devices, and in Figure 9.2a and b. Note that n- and p-channel depletion regions that junction depletion regions that change. In the MOSFET, they are the framework for the devices, and in Figure 9.2c and d. The MOSFETs have a heavy line before of a channel when no voltage on asymmetric geometry is used, terminals, and the source and drain

---

**Figure 9.3** MOSFET cross sections showing bias effects on the depletion regions: (a) drain voltage small, depletion region nearly uniform along the channel, (b) drain voltage large enough to cause significant variation in depletion–region thickness, (c) drain voltage exceeds the saturation value; channel extends only to $L' < L$. 

---

Charge-Control Analysis

As in the analysis of the JFET, we consider the MOSFET first at low drain-source voltages ($V_{DS}$ positive but small) so that the channel charge does not vary strongly with position between the source and drain electrodes. Under this condition the surface space-charge region is uniform along the channel (Figure 9.3a).

If both $V_S$ and $V_D$ are at zero volts and $V_G$ is increased from zero, the source and drain are initially isolated by a reverse-biased $p/n$ junction. When $V_G$ becomes larger than $V_{PB}$, a depletion region forms along the channel. The depletion region
9.1 Basic Theory

approximately

\[ I_{pd} = \mu_n C_{ox} \frac{W}{L} (V_G - V_T) V_{DS} \]  

Equation 9.1.4* agrees with intuition; for low drain bias the drain current varies linearly with \( V_{DS} \), and the conductance between the source and the drain \[ \mu_d (W/L) C_{ox} (V_G - V_T) \] is proportional to the effective gate bias \( (V_G - V_T) \).

If \( V_p \) is now increased until it is no longer negligible compared to \( V_G \), the analysis just given becomes inaccurate because the drain voltage \( V_p \) affects the channel bias near the drain and acts to reduce \( Q_n \), as described in Section 8.3. The space-charge region and channel charge for this situation are sketched in Figure 9.3b.

Distributed Analysis. Since the channel voltage varies continuously between the source and the drain, we require a differential equation to represent the channel charge exactly. The equation is derived by following the procedure used in Chapter 4 for analysis of the JFET. The incremental voltage drop along the channel is first found as a function of the channel current. Then, integration along a path extending from the source to the drain leads to an equation for \( I_D \) in terms of the applied voltages.

We assume the gradual-channel approximation to be valid, as we did in the JFET analysis in Chapter 4. This approximation states that fields in the direction of current flow are much smaller than are fields in the direction perpendicular to the silicon surface (mathematically that \( |\partial \phi / \partial y| \ll |\partial \phi / \partial x| \)). This assumption validates the use of a one-dimensional MOS analysis (as carried out in Chapter 8) to find the carrier concentrations and the depth of the depletion region under the channel.

For the present analysis, we also assume that the channel is appreciably longer than the depletion region at the drain (i.e., \( L' / L \approx 1 \)). In very short channel devices, this may not be true, and the channel charge can be influenced strongly by the drain voltage as well as by the gate voltage. This short-channel effect is considered further in Chapter 10. Finally, we assume that since drain current consists of electron flow through \( n \)-type regions, it is carried exclusively by the drift process.

Figure 9.4 shows an \( n \)-channel MOSFET that is biased at moderate drain voltage. An incremental length \( dy \) along the channel sustains a voltage drop \( dV_c \) that can be expressed (Problem 9.2) as the product of the drain current \( I_D \) and the incremental resistance \( dR \):

\[ dV_c = I_D dy = -\frac{I_D dy}{W \mu_n Q_a |y|} \]  

where \( Q_a (\ll 0) \) is a function of the channel voltage \( V_c(y) \), as well as the gate voltage \( V_G \). Separating variables and integrating Equation 9.1.5 from the source to the

* The subscript 0 on \( I_{pd} \) in Equation 9.1.4 is used to distinguish this zero-order formulation for MOSFET drain current. Other expressions for \( I_D \) and \( V_D \) will be subscripted in ascending order as they are introduced.
Figure 9.4  MOSFET cross section indicating the differential length \( dy \) along the channel. An ohmic voltage drop \( dV_c = I_D \, dy \) is sustained across \( dy \). The channel is \( W \) units wide.

\[
I_D = \frac{-\mu_n W}{L} \int_{V_S}^{V_D} Q_d(V_C) \, dV_C
\]  

(9.1.6)

The dependence of \( Q_n \) on channel voltage needed for the integral in Equation 9.1.6 was considered in Equation 8.3.16 and is repeated here for reference.

\[
Q_n = -C_{ox}(V_G - V_{FB} - 2|\phi_P| - V_C) + \sqrt{2\varepsilon_s q N_A(2|\phi_P| + V_C - V_B)}
\]  

(9.1.7)

As seen in Equation 9.1.7, channel charge depends on \( V_C \) through a linear term and also through a term embedded in the square root which represents depletion-layer charge \( Q_d \) (cf Equation 8.3.9). Equation 9.1.7 can be considerably simplified if \( Q_d \) is represented only by its value at the source and its variation along the channel is neglected. This approximation, which linearizes the equation, leads to a useful description of MOSFET behavior. Neglecting the variation of \( Q_d \) with \( V_C \), Equation 9.1.7 can be written

\[
Q_{n1}(V_C) = -C_{ox}(V_G - V_T - (V_C - V_S))
\]  

(9.1.8)

where, as in Equation 9.1.3, \( V_T \) refers to the threshold voltage at the source. Using Equation 9.1.8 in Equation 9.1.6, we obtain

\[
I_{D1} = \frac{\mu_n W}{L} \left( (V_G - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right)
\]  

(9.1.9)

where \( V_{DS} = V_D - V_S \). We refer to Equation 9.1.9 as the charge-control equation for drain current. The name “charge control” is applied here because Equation 9.1.9 can be obtained by considering the averaged charge-control expression for \( Q_n \). Equation 9.1.9 is clearly an approximation arising from the assumed form for the channel charge (Equation 9.1.8), but we shall see that it is still useful and frequently employed.

To study the MOSFET behavior curves of \( I_{D1} \) versus \( V_{DS} \) for the various slopes in the \( V_{DS} \) and \( V_G \) areas, and Equation 9.1.4 applies. Conductance diminishes, and the channel charge at \( V_{DS} = (V_G - V_T) \) lines for \( V_{DS} > (V_G - V_T) \) before the negative incremental conductive line to the channel breaks down can be seen by the channel \( (V_D - V_S) \) approximation. Equation 9.1.8 (applied to zero and then returned) indicates how the channel charge \( Q_{n1} \) changes as a function of the channel. The channel MOSFET behavior.

Although the theory is valid for the behavior of \( I_D \) and \( V_D \) that electrons in the channel move in the channel, they approach the values where they are accelerated up to the value \( V_{DS} \) before \( V_{DS} > (V_G - V_T) \) at which these electrons first-approximate analysis, this rate becomes constant or saturation curves of Figure 9.5a occur at \( V_{DS} \) at which \( Q_{n1} \) reaches a maximum.

\[
\text{Drain current} \quad \text{Drain voltage}
\]  

\((a)\)
To study the MOSFET behavior predicted by Equation 9.1.9, we have sketched curves of $I_D$ versus $V_{DS}$ for varying values of $V_G$ (> $V_T$) in Figure 9.5a. The curves have maximum slope in the vicinity of the origin where the conductance is greatest, and Equation 9.1.4 applies. For higher $V_G$, the channel charge and hence the conductance diminish, and the curves consist of a series of downward-facing parabolas having maxima at $V_{DS} = (V_G - V_T)$. In Figure 9.5a, the curves are drawn as dotted lines for $V_{DS} > (V_G - V_T)$ because they do not apply in this range of drain bias. The negative incremental conductance ($\partial I_D/\partial V_G$), indicated in this region, is an intuitive clue to their being physically unreasonable. The reason that our derivation breaks down can be seen by considering Equation 9.1.8 when the voltage along the channel ($V_D - V_S$) approaches and then exceeds ($V_G - V_T$). Under these conditions, Equation 9.1.8 (applied at the drain where $V_C = V_D$) predicts that $Q_{st}$ goes to zero and thereafter changes sign. This sequence is a physical impossibility because $Q_{st}$ represents an inversion-charge density that must always be negative for the $n$-channel MOSFET being analyzed.

Although the theory breaks down when $V_{DS} > (V_G - V_T)$, it is quite possible to deduce the behavior of $I_D$ versus $V_{DS}$ in this bias range. To do this, we recognize that electrons in the channel do not "see" a barrier as they approach the drain. On the contrary, they approach a high-field region with a very low electron density in which they are accelerated toward the drain (usually reaching limiting drift velocities) when $V_{DS} > (V_G - V_T)$. The drain current $I_D$ is thus determined by the rate at which these electrons arrive at the edge of the high-field region. In the first-order analysis, this rate is insensitive to $V_{DS}$, and the drain current thus becomes constant or saturates for $V_{DS} > (V_G - V_T)$. Hence, the maxima in the curves of Figure 9.5a occur at a drain voltage known as the saturation voltage $V_{DSat}$ at which $Q_{st} \rightarrow 0$ at the drain end of the channel.
\[ V_{\text{Dsat,1}} = (V_G - V_T) + V_S \]
\[ = V_G - V_{FB} - 2|\phi_p| - \frac{1}{C_{ex}} \sqrt{2q\varepsilon_s N_{d} [2|\phi_p| + (V_S - V_{th})]} \]  
(9.1.10)

The constant current for \( V_{DS} > V_{\text{Dsat}} \) is sketched in Figure 9.5b. In this range of bias, the end of the channel (i.e., the point at which \( Q_n \rightarrow 0 \)) is no longer at \( y = L \), but rather occurs at the point \( y = L' \) where the channel voltage \( V(y) = V_{\text{Dsat}} \). The current when \( V_D > V_{\text{Dsat}} \) (denoted \( I_{\text{Dsat,1}} \)) can therefore be calculated by using Equation 9.1.10 in Equation 9.1.9

\[ I_{\text{Dsat,1}} = \frac{\mu_n W C_{ex}}{2L} (V_G - V_T)^2 \]  
(9.1.11)

In Equation 9.1.11, \( V_T \) refers to the threshold voltage at the source—and depends, therefore, on \( V_S \). Since both Equations 9.1.11 and 9.1.9 are valid (within the assumptions made) at \( V_D = V_{\text{Dsat,1}} \), a plot of \( I_{\text{Dsat,1}} \) versus \( V_{\text{Dsat,1}} \) on the \( I_D \) versus \( V_D \) axes is an upward-facing parabola (dashed curve in Figure 9.5b). The overall \( I_D \) versus \( V_D \) curves (excluding breakdown effects) are thus plotted in Figure 9.5b by using Equation 9.1.9 for \( V_D < V_{\text{Dsat,1}} \) and Equation 9.1.11 for \( V_D > V_{\text{Dsat,1}} \).

The channel charge and depletion regions for \( V_D > V_{\text{Dsat,1}} \) are shown in Figure 9.3c. This figure emphasizes that \( L' \) decreases as \( V_D \) increases. This reduction in \( L' \) causes \( I_{\text{Dsat,1}} \) to increase slightly with increasing \( V_D \). The dependence on \( V_D \) arising from this effect can be incorporated by rewriting Equation 9.1.11 with \( L' \) in place of \( L \) and then expressing \( L' \) as a function of \( V_D \). This modification will be considered at a later point in the discussion.

**Water Analogy.** Physically, the operation of the MOSFET can be understood in terms of a water analogy. The free carriers correspond to water droplets. The source and drain are deep reservoirs whose relative elevation difference is analogous to the source-to-drain voltage difference. The channel region is like a canal with a depth that depends on the local value of the gate-to-channel voltage as sketched in Figure 9.6.

If the drain and source are held at the same potential, the water surface is level through the source, canal, and drain in our analogy (Figure 9.6a). When a drain-to-source voltage is imposed, the surface of the drain reservoir is lowered, causing a flow of water along the canal from the source to the drain. The flow speeds up as the elevation difference (analogous to \( V_{DS} \)) increases. Since the flow is continuous, the water velocity increases as the depth of water in the canal decreases toward the drain reservoir. At first, the flow through the canal depends both on its dimensions (as controlled by the gate) and on the elevation difference between the source and drain (Figure 9.6b). At the condition analogous to saturation, the flow becomes entirely limited by the flow capacity of the canal. After saturation, if the drain reservoir is lowered further, its surface becomes abruptly disconnected from the water surface at the drain end of the canal. In this condition, the flow into the drain resembles the free fall of water over a waterfall (Figure 9.6c). The rate of flow is equal to the delivery rate to the lip of the waterfall, independent of the total drop over the cataract [which is analogous to \( (V_D - V_{\text{Dsat}}) \)].

---

**Variable Depletion-charge Variation**

In order to derive an integral form of Equation 9.5.4 into Equation 9.1.6, let

\[ I_{D2} = \frac{\mu_n W}{L} \left\{ \frac{2}{5} \right\} \]

As noted previously, inversion charge exists sufficiently to cause \( Q_d \) to be automatically determined. Therefore, the corresponding differential equation is

\[ Q_d(L) = 0 = -C \]

---

**Figure**

Source water track (water channel)
\[ \phi_p + (V_S - V_B) \]  \hspace{1cm} (9.1.10)

Figure 9.5b. In this range of \( V_B \) (vertical) \( V_B \to 0 \) is no longer at \( y = L \), and the actual voltage \( V(y) \) \( V_B \) can be calculated using

\[ (9.1.11) \]

It depends not only on the source and drain but also on the \( V_B \) (horizontal) on y = 0. \( V_B \) is shown in Figure 9.5a. \( V_B \) increases. This reduction in \( V_B \) reduces the dependence on \( V_D \). The dependence on \( V_D \) can be found in Equation 9.1.11 with \( L' \) replacing \( L \) and \( V_B \). This modification will

9.1 Basic Theory

MOSFET can be understood as analogous to water droplets. The relative elevation difference is analogous to the voltage difference between gate and source. The flow of the channel region is like a canal flowing from source to drain. The flow speedup depends on water depth. The water depth in the channel can be varied by the gear and track (9.1.11). When the water is lower than the source, the flow decreases. When the source is lower than the drain, the flow increases. The water depth at the source determines the upper limit of the channel capacity. Lowering the source further increases the height of the waterfall at its edge.

Variable Depletion-Charge Analysis. It is not necessary to neglect the depletion-charge variation along the channel and approximate \( Q_d(V_c) \) by Equation 9.1.8, in order to derive an explicit equation for \( I_p \). If Equation 9.1.7 is inserted directly into Equation 9.1.6, the resulting integration has many terms, but is easily carried out (Problem 9.2). The result can be written

\[ I_p = \frac{W}{L} \left[ C_w \left( (V_D - V_B - 2|\phi_p| - \frac{1}{2} V_D - \frac{1}{2} V_S) \right) V_{DS} \right. \]

\[ \left. + \frac{2}{5} \sqrt{\varepsilon_s q N_a[(2|\phi_p| + V_D - V_B)^{3/2} - (2|\phi_p| + V_S - V_B)^{3/2}]} \right] \]  \hspace{1cm} (9.1.12)

As noted previously, the analysis leading to Equation 9.1.12 is valid as long as \( V_p \) is increased sufficiently to cause \( Q_p(L) \) to decrease to zero, the drain current saturates. To find the corresponding drain voltage \( V_{DSAT} \), therefore, we let \( V_c = V_D \) in Equation 9.1.7 and solve for \( V_{DSAT} \) by letting \( Q_p(L) \) equal zero.

\[ Q_p(L) = 0 = -C_w(V_D - V_{FB} - 2|\phi_p| - V_{DSAT}) + \sqrt{2\varepsilon_s q N_a(2|\phi_p| + V_{DSAT} - V_B)} \]  \hspace{1cm} (9.1.13)
From Equation 9.1.13, the saturation drain voltage is

\[ V_{Dsat} = V_G - V_{FB} - 2[\phi_p] - \frac{\varepsilon_r q N_a}{C_{ox}^2} \left( \sqrt{1 + \frac{2C_{ox}^2}{\varepsilon_r q N_a}} (V_G - V_{FB} - V_D) - 1 \right) \] (9.1.14)

As seen in Equation 9.1.14, the saturation drain voltage is independent of the source voltage because it is defined by a zero free-charge condition at the drain end of the channel.

To compare these results from the variable depletion-charge analysis with the equations derived earlier using charge-control analysis, we take corresponding voltages \( V_S = V_B = 0 \). Under this condition Equation 9.1.12 reduces to

\[ I_{D2} = \frac{\mu_n W}{L} \left\{ C_{ox} \left( V_G - V_{FB} - 2[\phi_p] - \frac{1}{2} V_D \right) V_D - \frac{2}{3} \sqrt{2 \varepsilon_r q N_a} \times \left[ (2[\phi_p] + V_D)^{3/2} - (2[\phi_p])^{3/2} \right] \right\} \] (9.1.15)

which can be compared to Equation 9.1.9 when the expression for \( V_I \) (Equation 8.3.18) is used in the equation. When that is done, Equation 9.1.9 takes the form

\[ I_{D1} = \frac{\mu_n W}{L} \left[ C_{ox} \left( V_G - V_{FB} - 2[\phi_p] - \frac{1}{2} V_D \right) V_D - 2V_D \sqrt{\varepsilon_r q N_a |\phi_p|} \right] \] (9.1.16)

Equation 9.1.16, the charge-control expression, can be thought of as a first-order representation for drain current because account has not been taken of the variation of depletion charge along the channel. Equation 9.1.15 treats the depletion-charge variation and is, therefore, a second-order representation. We shall see that although both representations are useful, often there are further effects to be considered.

---

**EXAMPLE Charge-Control and Variable Depletion-Charge Analyses**

Compare the output \( I_D \) versus \( V_G \) characteristics predicted by the charge-control equations with those obtained using the variable depletion-charge equations for the MOSFET. Consider a device for which \( \mu_n W/L = 1.2 \times 10^4 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1} \), \( C_{ox} = 3.98 \times 10^{-8} \text{ F cm}^{-2} \) (87 nm oxide), \( N_a = 2 \times 10^{11} \text{ cm}^{-3} \), and \( V_{FB} = -0.5 \text{ V} \).

**Solution**

The comparison can be made by using Equation 9.1.16 to represent the charge-control theory and Equation 9.1.15 for the variable depletion-charge analysis. As can be seen from the plot based on this comparison (Figure 9.7), the simpler charge-control equations (Equations 9.1.16 and 9.1.9) predict larger values for \( I_{Dsat} \) and \( V_{Dsat} \) for each value of gate voltage. The general behavior of the two sets of curves is, however, similar, and the currents predicted by both theories become the same as the drain voltage becomes roughly 20 times the drain voltage. Measurements of the MOSFET are presented in Figure 9.7.
From the graphical construction in Fig. 6.3c we expect that, as \( V_A \) approaches infinity, \( V_{DS} \) should approach \( V_{DS}^* \). This is easily checked to be the case with (6.2.17) (Prob. 6.3): As \( V_A \) approaches infinity, \( V_{DS} \) approaches the value of \( (V_{GS} - V_T) / \mu L \), which is, indeed, the value of \( V_{DS}^* \) for the model considered in the absence of channel length modulation [see (4.5.34)].

A method to combine the non-saturation and saturation expressions into a single expression using appropriate “smoothing functions” has been described elsewhere.\(^{40}\)

We emphasize again that (6.2.12) and (6.2.14) are good only for approximate predictions of the drain current in saturation. In applications where accurate prediction of the slope \( dl_{DS}/dV_{DS} \) is required (e.g., analog circuits), these equations are totally inadequate.

### 6.3 BARRIER LOWERING, TWO-DIMENSIONAL CHARGE SHARING, AND THRESHOLD VOLTAGE

#### 6.3.1 Introduction

The short- and narrow-channel effects to be considered in this section are of such a nature that they can be approximately described by the equations developed in Sec. 4.5.3, provided that in them the threshold voltage is replaced by another quantity, called the effective threshold voltage. This quantity will be seen to depend on channel length, channel width, source-substrate voltage, and drain-source voltage.\(^{39-106}\) We will initially consider the cases of short channels and of narrow channels separately.

#### 6.3.2 Short-Channel Devices

Let us review some of our assumptions for a long-channel device, shown in Fig. 6.4a; for simplicity, we assume here \( V_{DS} = 0 \). Neglecting the edge effects, as we have done so far, is equivalent to assuming that the situation between source and drain is the same as one would have with the source and drain removed, as shown in Fig. 6.4a, but with the channel somehow still communicating with the external world. Calculating \( Q_f \) and then \( I_{DS} \) by using this assumption provides accurate results as long as the value of \( L \) is large. Consider now a short-channel device, as shown in Fig. 6.4b, with the same process and biased with the same terminal voltages as the long-channel device. Here the edge effects practically extend over all of the channel. Neglecting these effects amounts to viewing the device as shown in Fig. 6.4d, which can hardly be expected to provide credible results. Indeed, assume a very small but nonzero \( V_{DS} \), so that drain current can flow without appreciably disturbing the picture in the channel in Fig. 6.4c. It has been found experimentally that the \( V_{GS} \) value required to produce a given \( I_{DS} \) value is smaller than what would have been expected from the picture in Fig. 6.4d, as shown in Fig. 6.5. To get some intuitive feeling for the reasons behind this, one can use a number of viewpoints found in the literature. One of these uses the concept of barrier lowering,\(^{60,66}\) referring to the following. The effect of decreasing \( L \) is to deplete more of the region under the inversion layer; this is easily visualized if the source and drain are imagined to approach one another as shown in
region un-

![Diagram of a MOS transistor with depletion and inversion regions labeled.](image)

**FIGURE 6.4**

(a) A long-channel transistor; (b) the channel of (a) with edge effects neglected; (c) a short-channel resistor; (d) the channel of (c) with edge effects neglected.

![Graph showing drain current vs. gate-source voltage with measured and expected curves.](image)

**FIGURE 6.5**

Drain current vs. gate-source voltage for very small, fixed drain-source voltage.

Fig. 6.4c. The deeper depletion region is accompanied by a larger surface potential,† which makes the channel more attractive for electrons. Thus the device can conduct more current than what would be predicted from long-channel theory for a given $V_{GS}$. Since, in long-channel theory, the drain current is an increasing function of $V_{GS} - V_T$, the long-channel equations may be made to describe artificially the current increase if $V_T$ is replaced by a smaller quantity $V'_T$, which will be called the effective threshold voltage. This quantity decreases as $L$ is decreased. For devices with the minimum possible channel length in a given technology, $V'_T$ can be smaller than $V_T$ by 50 to 200 mV. Notice that in the short-channel device of Fig. 6.4c the depletion

†In an energy band formulation, an increase in surface potential corresponds to a decrease in the potential energy for electrons. The potential energy "barrier" to the entrance of electrons in the channel is lowered; hence the name barrier lowering is used to describe these phenomena.
region under the channel can also be widened (and the corresponding surface potential can be increased) by raising the potential of the drain; this will further increase $I_{DS}$ above the value expected from other considerations. Hence, the effective threshold $V_T$ need also be made a decreasing function of $V_{DS}$. Unfortunately, it is not easy to come up with simple analytical results by using the barrier-lowering concept.

An alternative description goes as follows. The charge in the channel region is influenced by field lines emanating from all nearby structures. Two such structures considered in the analysis of long-channel devices are, of course, the gate and the substrate (“back gate”). In short-channel devices the source and drain are so close to all points in the ground that they can affect the latter through their proximity just as the gate does. In effect, the source and drain now play something of a gate role in addition to their normal function. Field lines emanating from all four structures (gate, bulk, source, and drain) and terminating on charges in the channel must be considered for an accurate description of the device. Essentially, charge control in the channel is shared by all four structures. Bringing the source and drain regions closer to all points in the channel is similar to bringing the gate closer to the channel. Since source and drain help support the channel charge, the latter effect increases as $L$ is decreased, for given $V_{GS}$ and $V_{SB}$. The corresponding increase in drain current, however, is not predicted by long-channel theory and is modeled instead by using an effective threshold as explained above. Also, increasing the drain potential increases the inversion layer charge just as increasing the gate potential would. The above picture again leads us to expect an effective threshold which decreases with decreasing $L$ and increasing $V_{DS}$.

This “charge-sharing” viewpoint has provided the basis for most of the analytical/empirical models for short-channel devices. We will expand on this idea below. But first we should warn the reader that charge-sharing models have been developed with the aim to obtain simple semiempirical expressions for describing very complex two-dimensional phenomena; in such developments, it has not been possible to justify rigorously all steps. To begin, let us assume that the equivalent interface charge $Q_n$ is zero for simplicity. In Figs. 6.4c and d, it can be imagined that one field line emanates from each positive charge $q$ on the gate. Such a field line terminates either on an electron in the inversion layer or on an ionized acceptor in the depletion region, neglecting the “fringing” field lines terminating on the $n^+$ regions. In Fig. 6.4d, all of the depletion region charges are “imaged” on gate charges through connecting field lines. In the more realistic picture of Fig. 6.4c, however, this is not the case. Some of the field lines terminating on ionized acceptor atoms near the $n^+$ regions can be originating from ionized donor atoms in the $n^+$ regions (inside a thin depletion layer there). Thus, only part of the depletion region charge is imaged on the gate charge in this case. If we can assume that the gate charge in $c$ and $d$ is the same, and since some gate charges in $c$ cannot be imaged on depletion charges, more of them are available to be imaged on inversion layer charges, which must, thus, increase to accept the extra field lines. The extra $I_{DS}$ observed is attributed to this extra inversion layer charge, although, strictly speaking, one would have to know the spatial distribution of that charge before such a conclusion could be reached. A number of arbitrary assumptions are obviously involved in the above arguments (Prob. 6.6). Some of these assumptions are critically considered elsewhere.75,85

Next, a large empirical step is taken: The short-channel device is assumed to behave as a fictitious device with a uniform depletion region, but with an effective
depletion region charge $\hat{Q}_B$, which has a smaller magnitude than the corresponding quantity $Q_B$ in Fig. 6.4d. Because $\hat{Q}_B$ and $Q_B$ are defined for devices with uniform depletion regions and equal gate areas, we will have $\hat{Q}_B/Q_B = \hat{Q}_B/Q_B$, where the primes denote quantities per unit area. The above assumption for an effective charge $\hat{Q}_B$ then implies that, instead of the long-channel threshold given by (3.4.14a),

$$V_T = V_{FB} + \phi_0 - \frac{Q_B}{C_{ox}}$$  \hspace{1cm} (6.3.1a)$$

$$= V_{FB} + \phi_0 + \gamma \sqrt{\phi_0 + V_{SB}}$$  \hspace{1cm} (6.3.1b)$$

we will have an effective threshold given by

$$\hat{V}_T = V_{FB} + \phi_0 - \frac{\hat{Q}_B}{C_{ox}}$$  \hspace{1cm} (6.3.2a)$$

or

$$\hat{V}_T = V_{FB} + \phi_0 + \frac{\hat{Q}_B}{Q_B} \gamma \sqrt{\phi_0 + V_{SB}}$$  \hspace{1cm} (6.3.2b)$$

We thus see that the above charge-sharing effect can be viewed as resulting in an effective decrease of the body effect coefficient by the factor $\hat{Q}_B/Q_B$. This may make sense intuitively since, in a short-channel device, the substrate’s control on the channel is reduced, as expected from Fig. 6.4c, and much of the channel is controlled by the gate, the source, and the drain. However, defining $(\hat{Q}_B/Q_B) \gamma$ as the new body-effect “coefficient” is not very practical, since $(\hat{Q}_B/Q_B)$ is itself a function of $V_{SB}$, as we will see. From the above equations we can write

$$\hat{V}_T = V_T + \Delta V_{TL}$$  \hspace{1cm} (6.3.3)$$

where $\Delta V_{TL}$ is the “threshold change” due to the short-channel effect, given by

$$\Delta V_{TL} = - \left(1 - \frac{\hat{Q}_B}{Q_B}\right) \gamma \sqrt{\phi_0 + V_{SB}}$$  \hspace{1cm} (6.3.4)$$

With $|\hat{Q}_B| < |Q_B|$, $\Delta V_{TL}$ is a negative number as expected from the above discussion.

The meaning of $\Delta V_{TL}$ is illustrated in Fig. 6.6. Note that the $V_{SB}$ dependence of $\Delta V_{TL}$ in (6.3.4) does not include the total effect of $V_{SB}$ on $V_T$. For a given $V_{SB}$, one must first find $V_T$ for a long-channel device and then change it by the corresponding $\Delta V_{TL}$ to arrive at the value of $\hat{V}_T$. The value of $\hat{V}_T$ obtained in this way is widely used in place of $V_T$ in long-channel equations, such as (4.5.37). There is a large amount of literature on how $\hat{V}_T$ should be calculated, but very little on why it is justifiable to use this value in long-channel equations for the drain current (derived by using the
gradual-channel approximation) and expect the result to be correct for short-channel devices. The issue is not just what value of threshold voltage should be used in such equations when modeling short-channel devices, but rather why the very form of such equations should be valid for short-channel devices in the first place. A convincing argument in this regard is not apparent. However, satisfactory agreement with measurement has led to wide use of the aforementioned approach.

Of the many techniques proposed for determining the quantity $Q_R/Q_R$, let us consider one as an example. The $n^+$-region edge will be assumed cylindrical with radius equal to the junction depth $d_j$, as shown in Fig. 6.7a. Next, the depletion regions around each of them and below the inversion layer are drawn as if each existed by itself (i.e., no interaction is assumed), and then they are joined together. The width of all three regions is considered equal by assuming that the junction built-in potential $\phi_{bi}$ is equal to $\phi_{nj}$; thus, this width is, from (3.4.8),

$$d_R = \zeta \sqrt{\phi_{bi} + V_{SB}}$$

(6.3.5a)

with

$$\zeta = \sqrt{\frac{2\epsilon_s}{qN_A}}$$

(6.3.5b)

In this picture, $L$ is assumed to be at least large enough so that a region of trapezoidal cross section can be defined as shown in Fig. 6.7a. $Q_R$ is taken to be equal to the charge in that region, and $Q_n$ is assumed to be the charge corresponding to a rectangle of the same depth and length $L$. Simple geometry then gives (Prob. 6.7):

$$\frac{\hat{Q}_n}{Q_R} = 1 - \frac{d_j}{L} \left( \sqrt{1 + \frac{2d_n}{d_j}} - 1 \right)$$

(6.3.6)

$\hat{V}_T$ and $\Delta V_T$, can now be found from (6.3.2b) and (6.3.4) to (6.3.6). If $L$ is too small, the trapezoid in Fig. 6.7a becomes a triangle; this case is considered in Prob. 6.8. In the following, we assume that $d_n < L/2$, so that this case does not occur.
FIGURE 6.7
(a) Trapezoid approximation used for deriving the effective bulk charge; (b) the limiting case of (a) for very deep $n^+$ regions.

For some practical applications the above formulation is too complicated. It has been suggested that for approximate calculations the quantity in parentheses in (6.3.6) can be approximated by using a Taylor expansion and neglecting the higher-order terms, resulting in (Prob. 6.7):

$$\frac{\hat{Q}_B}{Q_B} = 1 - \frac{d_B}{L}$$  \hspace{1cm} (6.3.7)

This expansion will be more accurate if $d_B/L$ is small. Negligible $d_B/L$ corresponds to Fig. 6.7b, for which (6.3.7) can easily be obtained directly. Since, for larger values of $d_B/L$, the expression can be in considerable error, one can help expand its region of validity by introducing a parameter:

$$\frac{\hat{Q}_B}{Q_B} = 1 - \beta \frac{d_B}{L}$$  \hspace{1cm} (6.3.8)

where nominally $\beta = 1$, but one can allow for empirically adjusting this parameter to obtain best overall fit in a given region.† Using this equation and (6.3.5) in (6.3.2b), we obtain

$$\hat{V}_T = V_{FB} + \phi_0 + \gamma \sqrt{\phi_0 + V_{SB}} \left(1 - \frac{\beta \zeta}{L} \sqrt{\phi_0 + V_{SB}}\right)^2$$ \hspace{1cm} (6.3.9)

†The same can be done with $\phi_0$ in the equations in this section. Recall that the value of this parameter was somewhat ambiguous even for long-channel devices (see Secs. 2.5.2 and 4.5.1). Thus, although the value of $2\phi_0$ is widely used for $\phi_0$, better results can be obtained if one allows $\phi_0$ to be adjusted empirically. Given the heavy empirical nature of the charge-sharing model we are discussing, this is reasonable.

‡It would appear from this equation that, with a sufficiently large $V_{SB}$, $dV_T/dV_{SB}$ can become negative. However, it can be easily checked by differentiation that, for this to happen, $V_{SB}$ would need to be so large that $d_B > L/2$. In that case, though, the development of this equation is not valid [see after (6.3.6)]. This is a good example of the wrong conclusions that can be reached by blindly using an equation and forgetting the assumptions that led to its development.

FIGURE 6.8
Effect of trapezoid approximation. The body of the trapezoid can be approximated by short columns of trapezoidal shapes. Negligible $d_B/L$ diminishes the error. The region of interest of the trapezoid is $d_B - L$. A better approximation would be to neglect the shortening of $d_B$. If $d_B$ is large, this approximation is valid. With $C$, the effective area under the column $d_B$ can be approximated by $C$. If $d_B$ is small, the error is negligible.
The factor in parentheses can be thought of as an effective reduction in the body effect coefficient. As expected, this reduction is more severe for smaller $L$. The behavior predicted by the above equation is illustrated in Fig. 6.8. It is seen that, for short channels and large $V_{SB}$, the dependence of the effective threshold on $V_{SB}$ diminishes. This is because the factor in parentheses in (6.3.9) becomes very small. Physically, this corresponds to the fact that the lower base of the trapezoid in Fig. 6.7 diminishes in length. The bottom of the trapezoid is then practically cut off from the rest of the substrate, and thus, the control of the substrate on the charge inside the trapezoid is small (Prob. 6.8).

A calculation of the reduction $\Delta V_T$ corresponding to the above effects is illuminating. Using (6.3.5) and (6.3.8) in (6.3.4), and recalling that $\gamma = (2q\varepsilon_s N_A)^{1/2}/C'_{ox}$, with $C'_{ox} = \varepsilon_{ox}/t_{ox}$, we obtain

$$\Delta V_{IL} = -2\beta_1 \frac{\varepsilon_s}{\varepsilon_{ox}} \frac{t_{ox}}{L} (\phi_0 + V_{SN})$$

(6.3.10)

From this equation it is clear that the channel length and the oxide thickness have a competing role. Although decreasing $L$ tends to increase the short-channel effect, decreasing $t_{ox}$ tends to decrease it. This is because then the gate is closer to the channel and is thus better able to keep control of the depletion region charge (as opposed to releasing this control to the other structures surrounding the channel).

Good agreement has been reported between (6.3.9) and (6.3.10) and experimental results, with $\beta_1 = 1.69$. However, other work has produced models equivalent to taking $\beta_1$ proportional to a negative power of $L$, or even replacing $\beta_1/L$ by a negative exponential in $L$.

**EFFECT OF DRAIN-SOURCE VOLTAGE.** The above results have been derived for negligible $V_{DS}$. For fixed $V_{SB}$, if $V_{DS}$ (and thus $V_{DB}$) is increased, the depletion region
width around the drain will widen. This, in turn, following the lines of the above arguments, will decrease $|\hat{Q}_B|$ below the values obtained so far, thus further decreasing $\hat{V}_T$. Thus, for short-channel devices, $\hat{V}_T$ becomes a decreasing function of $V_{DS}$. The effective charge $\hat{Q}_B$ is still calculated using a method similar to the one illustrated in Fig. 6.7a, only now the trapezoid will become distorted. The resulting expression is considered in Prob. 6.11. If $d_J$ is large and certain simplifications are made (Prob. 6.11), we can obtain

$$\frac{\hat{Q}_B}{Q_B} = 1 - \beta_1 \frac{d_{BS} + d_{BD}}{L}$$

(6.3.11)

where $d_{BS}$ and $d_{BD}$ are the depletion region widths of the source and the drain, respectively, and $\beta_1$ is again a constant, nominally equal to 1 but which can be adjusted for better fit to experimental results. Note that for $V_{DS} = 0$ ($d_{BS} = d_{BD} = d_B$), the above equation reduces to (6.3.8). The quantity $(d_{BS} + d_{BD})/2$ is, from (6.3.5a) and a similar equation for the drain depletion region,

$$\frac{d_{BS} + d_{BD}}{2} = \frac{\xi}{2} \left( \sqrt{\phi_0 + V_{SB}} + \sqrt{\phi_0 + V_{DB}} \right)$$

(6.3.12)

We can write $V_{DB} = V_{SB} + V_{DS}$. For small $V_{DS}$, the second square root can be expanded in a Taylor series around $V_{DS} = 0$ and approximated by the first two terms; this results in (Prob. 6.11):

$$\frac{d_{BS} + d_{BD}}{2} = \xi \left( \sqrt{\phi_0 + V_{SB}} + \frac{\beta_2 V_{DS}}{\sqrt{\phi_0 + V_{SB}}} \right)$$

(6.3.13)

where $\beta_2 = 0.25$ results from the expansion. To increase the region of validity of this expression, however, empirical adjustment of $\beta_2$ may be needed. Using now (6.3.13), (6.3.11), and (6.3.2b), we obtain,

$$\hat{V}_T = V_{FB} + \phi_0 + \gamma \sqrt{\phi_0 + V_{SB}} \left[ 1 - \frac{\beta_2 \xi}{L} \left( \sqrt{\phi_0 + V_{SB}} + \frac{\beta_2 V_{DS}}{\sqrt{\phi_0 + V_{SB}}} \right) \right]$$

(6.3.14)

and thus

$$\Delta V_{TL} = -2 \beta_1 \frac{\xi}{\epsilon_{ox}} \frac{L}{\epsilon_{ox}} \left[ (\phi_0 + V_{SB}) + \beta_2 V_{DS} \right]$$

(6.3.15)

The behavior predicted by the above equations is shown in Fig. 6.9.

Although (6.3.15) was derived by using the charge-sharing idea, it predicts a variation of $V_T$ with $V_{DS}$ in the same direction as considerations using the barrier lowering concept discussed in the beginning of this section. The name drain...
1) **Triangular Quantum Well**

\[ V(3) = q \sqrt{F_s^2} \]

\[ \Rightarrow E_c = \left( \frac{\hbar^2}{2m} \right)^{1/3} \left[ 3qF_s \sqrt{1 + \frac{3}{4}} \right] \]

2) **Background Doping is Low**

\[ \Rightarrow E_F = \frac{q\sqrt{N_s}}{\varepsilon} \]

Substituting \( F_s = \frac{q\sqrt{N_s}}{\varepsilon} \) in the expression for \( E_c \) gives:

\[ E_0 = \frac{\varepsilon}{2} N_s^{2/3} \]

\[ E_1 = \frac{\varepsilon}{2} N_s^{2/3} \]
For only two subbands:

\[ n_s = \frac{D}{qV} \sum_{i=0}^{qV} \ln \left( 1 + \exp \left( \frac{q(V_i - E_f)}{k_B T} \right) \right) \]

where \[ D = \frac{\sqrt{m^*}}{11 \hbar^2} \]

(= 3.24 x 10^17 m^2/V.s for GaAs)

\[ n_s = \frac{D}{qV} \ln \left[ \left( 1 + \exp \left( \frac{qV}{k_B T} [V_d^+ - E] \right) \right) \right] \]

\[ \times \left( 1 + \exp \left( \frac{qV}{k_B T} [V_d^- - E] \right) \right) \]

where

\[ V_d^+ \]

Fraction of \( E_c \) supported by

\[ V_d^- \]

Voltage in GaAs

\[ V_d \]

Fraction of \( E_c \) supported by voltage in AlGaAs
From \( N_{so} = f(V_{di}^+) \)

We obtain \( V_{di}^+ \) which is the Fermi level position with respect to the bottom of the conduction band.

Therefore

\[
\exp \left( -\frac{N_{so}}{DVT} \right) = \left( 1 + \exp \left\{ \frac{-qE_0}{k_B T} \right\} \right) \left( 1 + \exp \left\{ \frac{-qE_1}{k_B T} \right\} \right)
\]

where \( V_T = \frac{k_B T}{q} \)

Setting \( \exp \left( \frac{V_{di}^+}{V_T} \right) = x \) we get

\[
\exp \left( -\frac{N_{so}}{DVT} \right) = \left( 1 + \frac{x}{a_0} \right) \left( 1 + \frac{x}{a_1} \right)
\]

or

\[
x^2 + x(a_0 + a_1) - a_0 a_1 (1 - \exp \left( -\frac{N_{so}}{DVT} \right)) = 0
\]
RUTHLESS APPROXIMATIONS

\[
\delta_0 = 2.5 \times 10^{-12} \text{ eV cm}^{4/3}
\]
\[
\delta_1 = 3.2 \times 10^{-12} \text{ eV cm}^{4/3}
\]

\[
\text{OR Asume that } E_0 \text{ & } E_1 \text{ are not separable in energy}
\]
\[
\delta_0 = \delta_1
\]

\[
\text{OR } a_0 = a_1
\]

Linearizing the exponentials

\[
1 + \frac{V_{di}^+}{V_t} = \frac{2A}{2} + \left[ A^2 - A \left( 1 - \left( 1 + \frac{n_s}{2D V_t} \right) \right) \right]^{1/2}
\]
\[
= -A + A \left[ 1 + \frac{n_s}{2DV_t} \right]
\]
\[
= + \frac{An_s}{2DV_t}
\]

\[
V_{di}^+ = \frac{An_s}{2D} - V_t = \Delta E_f(1) + an_s
\]
Charge Density at Heterojunction (AlGaAs - GaAs),

Assume: \( z < 0 \quad V = \infty \)
\( z > 0 \quad V = qF_Sz \) Triangular Potential Well

\( N_D, N_A = 0 \)

\[
E_i = \left( \frac{\hbar^2}{2m} \right)^{1/3} \left[ 3qF_S\pi \left( \frac{i + 3}{4} \right)^2 \right]^{2/3}
\]

\( \epsilon F_S = q n_s \quad n_s = \sum_{i=0}^{\infty} n_i \)

\( E_0 = \gamma_0 (n_s)^{2/3} \)
\( E_1 = \gamma_1 (n_s)^{2/3} \)

\( \gamma_0 = 2.5 \times 10^{-12} \text{ Vm}^{4/3} \) (SDH Data)
\( \gamma_1 = 3.2 \times 10^{-12} \text{ Vm}^{4/3} \)

\[
n_s = \frac{Dk_BT}{q} \sum_{i=0}^{1} \ln \left( 1 + \exp \left[ q \left( E_F - E_i \right) / k_BT \right] \right)
\]

\[
D = \frac{q\epsilon}{\pi \hbar^2} = 3.24 \times 10^{17} \text{ m}^{-2} \text{ V}^{-1} \text{ (SDH Data)}
\]
Charge Density vs. $E_F$ (1, AlGaAs)

Poisson's Equation

$$\frac{d^2 V}{dx^2} = \frac{\rho(x)}{\epsilon}$$

$$\rho(x) = q \left[ N_d^+ (x) - n(x) \right]$$

$$N_d^+ (x) = \frac{N_d}{1 + g \exp \left( \frac{E_d + qV}{k_B T} \right)}$$

$$n(V) = \frac{N_c \cdot \exp \left( qV/k_B T \right)}{1 + 1/4 \exp \left( qV/k_B T \right)} \left[ E_F = 0 \right]$$

Boundary Conditions

$F (-\infty) = 0$, $\rho (-\infty) = 0$

$$n_{so} = \frac{\epsilon_s}{q} F(d_i^-) = \frac{\epsilon_s}{q} F(0)$$

$$n_{so} = \left\{ \frac{2 \epsilon N_d}{q} \left[ -V(d_i^-) + V(-\infty) + \delta \right] + N_d^2 d_i^2 \right\}^{1/2} - N_d d_i$$

$p$ accounts for F-D statistics in degeneracy

$$\delta = -\frac{k_B T}{q} \left\{ \ln \left[ 1 + \exp \left( \frac{E_d + V(-\infty)}{k_B T} \right) \right] + 4 \frac{N_c}{N_d} \ln \left[ 1 + 1/4 \exp \frac{V(-\infty)}{k_B T} \right] \right\}$$
Charge Density vs. $E_F$ (2, GaAs)

TWO SUB-BANDS

$$n_{so} = D \frac{k_B T}{q} \ln \left[ \left( 1 + \exp \left\{ \frac{q}{k_B T} \left[ V(d_i^+ - E_0) \right] \right\} \right) x \left( 1 + \exp \left\{ \frac{q}{k_B T} \left[ V(d_i^- - E_1) \right] \right\} \right) \right]$$

$$E_o = \gamma_o n_{so}^{2/3} : E_1 = \gamma_1 n_{so}^{2/3}$$

$$V(d_i^+) = \Delta E_c + V(d_i^-)$$

SIMPLIFY: $V(d_i^+) = \Delta E_{FO}(T) + a n_{so}$

$$a = 0.125 \times 10^{-16} \text{ V m}^{-2}$$

$$\Delta E_{FO} = 0 \text{ eV } @ 300K: 0.025 \text{ eV } @ \lesssim 77K$$

SINCE: $n_{so}$(AlGaAs) = $n_{so}$(GaAs)

$$n_{so} = \left( \frac{2N_d \varepsilon}{q} \left[ \Delta E_c - E_{FO}(T) + \delta + V(-\infty) \right] \right)^{1/2} - N_d \left( d_i + \Delta d \right)$$

where $\Delta d = \frac{\varepsilon \phi}{q} = 80 \text{ Å}$
Charge Control of Two Dimensional Electron Gas

\[ n_s \quad \text{at} \quad qd = \left[ V_g - \left( \frac{\epsilon}{\gamma_b} V_{p2} + V(d_i^+) - \Delta E_c \right) \right] \]

\[ V_{p2} = \frac{qN_qd_d^2}{2\varepsilon} \]

From Density of States and F-D Statistics

\[ \exp \left[ -\frac{V(d_i^+)}{V_T} \right] = \frac{a_o + a_1}{2} + \left( \frac{a_o + a_1}{2} \right)^2 \]

\[ -a_0 a_1 \left[ 1 - \exp \left( -\frac{n_s}{D V_T} \right) \right]^{1/2} \]

\[ a_o = \exp \left( \frac{E_o}{V_{T_i}} \right) \quad a_1 = \exp \left( \frac{E_i}{V_T} \right) \quad V_T = \frac{k_B T}{q} \]

\[ V(d_i^+) = \Delta E_{Fo}(T) + an_s \quad \text{(from figure)} \]

\[ n_s = \frac{\varepsilon}{q(d + \Delta d)} \left( V_g - V_{off} \right) \]

\[ V_{off} = \theta_b - \Delta E_c - V_{p2} + \Delta E_{Fo} \]

\[ \Delta d = \frac{\varepsilon a}{q} = 80 \text{ Å} \]

\[ \Delta d = \text{Distance of centroid of charge distribution from heterointerface} \]
Current Voltage Characteristics of MODFETs (Gradual Channel Approximation)

\[ n_s(x) = \frac{\varepsilon}{d + \Delta d} [V_g - V_{off} - V(x)] \]

\( V(x) \) = Channel Potential

\[ I_{ds} = q n_s v(F) Z \]

If Velocity Saturation Occurs at Drain Edge of Gate

\[ I_{dss} = Q (C \text{ cm}^{-2}) \cdot V_s \]

\[ I_{dss} = \frac{\varepsilon Z}{d + \Delta d} (V'_g - V_0) \]

\( V'_g = V_g - V_{off} \)

\( V_0 = F_s L \)

\[ (g_m)_{max} = \frac{q\mu n_{so}}{L \left( 1 + \left| \frac{q\mu n_{so} (d + \Delta d)}{\varepsilon V_s L} \right|^2 \right)^{1/2}} \]
Parasitic MESFET ($n_3, \mu_3$) in AlGaAs

- Transconductance Compression
- Gain Compression
Types of Modulatic Doped Structures, 1

GaAs (n⁺, u)

AlGaAs (n⁺)

GaAs

(a) UNIFORM DOPING [EASIEST TO GROW]

GaAs

AlGaAs (u)

GaAs n⁺

GaAs

(b) SUPERLATTICE DOPING (Δᵇ > Δᵃ) [REDUCED LIGHT SENSITIVITY]
Types of Modulation: Vopeda Structure

GaAs (n⁺, u)
AlGaAs (u)
δ(n⁺)
GaAs

(c) DELTA DOPING
MAXIMUM CHARGE TRANSFER
EXCELLENT CHARGE CONTROL
LOW GATE LEAKAGE
REDUCED LIGHT SENSITIVITY

GaAs (n⁺, u)
AlGaAs (u)
δ(n⁺)
GaAs

(d) QUANTUM WELL
[GOOD CARRIER CONFINEMENT]
Types of Modulation Doped Structures, 3

p⁺ GaAs
n AlGaAs
u GaAs
(e) p⁺ GATE
ENHANCES GATE BARRIER

GaAs (u)
AlGaAs (p)
In₀.₁₅Ga₀.₈₅As (u)
GaAs

(f) p - MODFET [ALLOWS COMPLEMENTARY LOGIC]
Current Voltag Characteristics of MODFETs (Short Gates)

\[
(g_m)_{\text{max}} = \frac{q \mu n_{SO}}{L \left(1 + \left[\frac{q \mu n_{SO} (d + \Delta d)/\varepsilon V_{S} L}{\varepsilon} \right]^2\right)^{1/2}}
\]

\[
d = d_d + d_i
\]

\[
(V_{Po})_{2D} = \frac{q(d + \Delta d)n_{SO}}{\varepsilon}
\]

\[
d_d = \left[2 \left(\phi_b - \Delta E_c - V_{off}\right) \frac{\varepsilon}{qN_D}\right]^{1/2}
\]

\[
A_s N_D \cdot d_d^4 - (g_m)_{\text{max}}^4
\]

\[
(g_m)_{\text{max}} = \frac{\varepsilon \mu (V_{Po})_{2D}}{(d + \Delta d) L \left\{1 + \left[\frac{\mu (V_{Po})_{2D}/V_{S} L}{\varepsilon} \right]^2\right\}^{1/2}}
\]

\[
A_s L \frac{\mu (V_{Po})_{2D}}{V_{S} L} \gg 1
\]

\[
(g_m)_{\text{short}} = \frac{\varepsilon V_{S}}{(d + \Delta d)} = C_g \cdot V_{S}
\]
DEVICE REQUIREMENTS
FOR mm-WAVE APPLICATIONS

- **HIGH CURRENT GAIN** \((f_t \uparrow)\)
  \[
  f_T = \frac{g_m}{2\pi C_{gs}} = \frac{V_e}{2\pi L_{eff}}
  \]
  \(\*V_e \uparrow, L_{eff} \uparrow\)

- **HIGH VOLTAGE GAIN**
  \[
  A_V = \frac{g_m}{g_o}
  \]
  \(g_m: TRANSCONDUCTANCE, g_o: OUTPUT CONDUCTANCE\)
  \(\*I_{SUB} \) (PARASITIC CONTRIBUTION TO \(g_o\))

- **HIGH POWER GAIN** \((f_{\max})\)
  \[
  f_{\max} = (f_T/2)[(R_g + R_i + R_s) g_o + 2\pi f_T R_g C_{dg}]^{-\frac{1}{2}}
  \]
  \(\*f_T \uparrow, R_g \uparrow, g_o \uparrow\)

- **HIGH DRIVE CAPABILITY (FAN OUT)\)**
  \(\*n_s \uparrow\)

AUGUST 1988
Scaling issues - Polarity and Channel Modulation

- Good channel modulation ($g_m I$)
- Effective gate shielding of drain field ($g_0 I$)

- Poor channel modulation ($g_m I$)
- Weak gate shielding of drain field ($g_0 I$)

- For Reduced Short Channel Effects $\frac{L_g}{a} > 5$

- For $L_g = 0.1 \mu m$ $a \leq 200 \text{ Å}$

- $N_D^{\text{min}} = \frac{2 \varepsilon_S \Phi}{qd^2} > 2 \times 10^{18} \text{ cm}^{-3}$
NATURE OF DEVICE CURRENT WITH VARIOUS BUFFER LAYERS (VEHICLE GaAs MESFET)

1. NOMINALLY UNDOPED (p-) BUFFER

2. p-TYPE GaAs BUFFER

3. AlGaAs BUFFER

(1) GaAs BUFFER: (2) p GaAs (3) AlGaAs

17448-1R1

AUGUST 1988
IMPORTANCE OF GATE RECESS IN
mm-WAVE FETs

- $g_{m_0} \propto C_g \cdot V_S$
- $g_m = \frac{g_{m_0}}{1 + g_{m_0} R_S}$
- $R_S = R_C + R_S(\rho_{S_1}) + R_S(\rho_{S_2})$
- $\rho_{S_1} \sim 100 \ \Omega/\square ; \rho_{S_2} \sim 1 \ \text{k}\Omega/\square$
- $L_g^* = L_g + L_{RD}$
- $\frac{L_g^*}{a} > \frac{L_g}{a} (g_{o1})$
- $L_g^* > L_g (f_T/4)$

- EXTREMELY IMPORTANT TO MINIMIZE $L_{RS}$
ADVANTAGES OF Ga.47In.53As TRANSISTORS

- HIGH ELECTRON MOBILITY
  * $R_S \downarrow$, $R_D \downarrow$

- LARGE $\Gamma$-$L$ VALLEY SEPARATION

- HIGH PEAK VELOCITY
  * $g_m \uparrow$, $f_T \uparrow$

- STRONGER VELOCITY OVERSSESOT
  * $g_m \uparrow$, $f_T \uparrow$

\[
\begin{align*}
\mu (\text{Ga.47In.53As}) &= 13,000 \text{ cm}^2\text{V}^{-1}\text{S}^{-1} \\
\mu (\text{GaAs}) &= 8,000 \text{ cm}^2\text{V}^{-1}\text{S}^{-1} \\
\Delta \Gamma-L (\text{Ga.47In.53As}) &= 0.55\text{eV} \\
\Delta \Gamma-L (\text{GaAs}) &= 0.31\text{eV} \\
V_p (\text{Ga.47In.53As}) &= 2.7 \times 10^7 \text{cms}^{-1} \\
V_p (\text{GaAs}) &= 2 \times 10^7 \text{cms}^{-1}
\end{align*}
\]
HIGH ELECTRON MOBILITY TRANSISTORS

ADVANTAGE OF MODULATION DOPING

- HIGH MOBILITY BECAUSE OF REDUCED COULOMBIC SCATTERING
- HIGH VELOCITY
- STRONGER VELOCITY OVERSHOOT

n-TYPE SEMICONDUCTOR
UNIFORM DOPING

SELECTIVELY DOPED SEMICONDUCTOR
MODULATION DOPING

CAUTION:
PICTORIAL REPRESENTATION ONLY

AUGUST 1988
III-V HEMT STRUCTURES

<table>
<thead>
<tr>
<th>ALGaAs HEMT</th>
<th>N⁺GaAs</th>
<th>N⁺Al₃Ga₇As</th>
<th>P⁺GaAs</th>
<th>P⁺GaAs</th>
<th>Si GaAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOURCE</td>
<td>CAP</td>
<td>GATE</td>
<td>DRAIN</td>
<td>CAP</td>
<td>SUBSTRATE</td>
</tr>
<tr>
<td>Pseudomorphic HEMT</td>
<td>N⁺GaAs</td>
<td>N Al₁₅Ga₈₅As</td>
<td>P⁺In₁₅Ga₈₅As</td>
<td>P⁺GaAs</td>
<td>Si GaAs</td>
</tr>
<tr>
<td>InGaAs HEMT</td>
<td>N⁺In₅₃Ga₄₇As</td>
<td>N Al₄₈In₅₂As</td>
<td>N⁺In₅₃Ga₄₇As</td>
<td>N⁺Al₄₈In₅₂As</td>
<td>Si InP</td>
</tr>
</tbody>
</table>

AUGUST 1998
WHAT ABOUT VELOCITY OVERSHOOT?

- NOT STRONG IN HIGHLY DOPED CHANNELS

- MORE PROBABLE IN MATERIALS SYSTEMS WITH
  LOW ELECTRON EFFECTIVE MASS
  HIGH ELECTRON MOBILITY
  HIGH PEAK VELOCITY
  (Ga_{0.47}In_{0.53}As)

UNDOPED CHANNELS – Al_{0.48}In_{0.52}As –
Ga_{0.47}In_{0.53}As HEMTS
COMPARISON OF AlGaAs–GaAs AND AllInAs–GaInAs HETEROSTRUCTURES

AlGaAs–GaAs
- $\Delta E_C = 0.25 \text{ eV}$
- $n_s = 1 \times 10^{12} \text{ cm}^{-2}$
- $\mu = 7000 \text{ cm}^2 \text{ V}^{-1} \text{s}^{-1}$
- $v_s = (1-2) \times 10^7 \text{ cms}^{-1}$
- $v_p (0.25 \mu\text{m}) = (3-4) \times 10^7 \text{ cms}^{-1}$

AllInAs–GaInAs
- $\Delta E_C = 0.55 \text{ eV}$
- $n_s > 3 \times 10^{12} \text{ cm}^{-2}$
- $\mu = 10,000 \text{ cm}^2 \text{ V}^{-1} \text{s}^{-1}$
- $v_s = (1-2) \times 10^7 \text{ cms}^{-1}$
- $v_p (0.25 \mu\text{m}) = (6-7) \times 10^7 \text{ cms}^{-1}$
Chapter 2
Design of mm-wave JHEMTs

A cross section of the p⁺-GaInAs/n-AlInAs/GaInAs JHEMT is shown in Figure 2.1. The device may be divided into two regions: the gate region (immediately under the gate metal) and the access regions (between the gate and the ohmic contacts). The two regions have different surface potentials as shown in the associated band diagrams. To achieve a high performance device, both regions must be effectively designed.

The design of mm-Wave JHEMTs begins with a derivation of the Lever Rule which governs the transfer of electrons from the donor layer to the channel. Second, the design criteria pertaining to the access regions are discussed. Next, a one-dimensional threshold voltage model is fully developed. Afterwards, a lumped-element approximation for the input impedance of the JHEMT is derived. Finally, the five design philosophies of the gate region are presented.

2.1 The Lever Rule Layer Design Model

The Lever Rule model\(^1\) dictates the distribution of electrons in a modulation-doped heterostructure. Given knowledge about the (δ-doped) donor layer, the model provides guidance when choosing the thicknesses of the spacer layer and barrier layer required to achieve a certain electron concentration in the channel. The model assumes
Figure 2.1. Cross-section of the p⁺-GaInAs/n-AllInAs/GaInAs JHEMT defining the gate and access regions. The associated band diagrams of these regions are also shown.
only that the donor impurities are fully ionized.

The modulation-doped band structure for the model is given in Figure 2.2. The thicknesses $d_2$ and $d_1$ are the thicknesses of the barrier and spacer layers, respectively. The sheet electron concentrations $n_d$ and $n_s$ are electron concentrations in the donor and channel layer, respectively. $\phi_b$ is the Schottky barrier height on the wide-bandgap material (e.g. AlInAs), and $\Delta E_c$ is the conduction band discontinuity at the heterojunction as defined in Chapter 1. The value $E_{rx}$ is the distance from the conduction band edge at the hetero-interface to the Fermi level as shown in the figure. Lee and coworkers\(^2\) have shown that $E_{rx}$ may be given by the following linear approximation\(^3\):

$$E_{rx} = \Delta E_{ro}(T) + a \cdot n_s \quad \text{[2.1]}$$

where:

$\Delta E_{ro}(T)$ is the zero-intercept of the linearized $E_{rx}(n_s)$. 

23
Chapter 2

function. note: $\Delta E_{r0}(300K) = 0$

$a$ is the slope of the linearized $E_{rx}(n_i)$ function.

$n_i$ is the 2-DEG concentration.

The rest of the analysis comes from applying Poisson's equation and Gauss' Law to the band structure. The electric field, $E_1$, may be written as:

$$E_1 = \frac{q \cdot n_i}{\varepsilon_2}$$  \[2.2\]

where $\varepsilon_2$ is the dielectric constant in the large bandgap material (e.g. AllInAs). At $T=300K$, the potentials in the structure at $300K$ may be summed to obtain the following equation which may be solved exactly:

$$-V_s + \frac{\Phi_s}{q} - E_2 \cdot d_2 + E_1 \cdot d_1 - \frac{\Delta E_s}{q} + an_i = 0$$  \[2.3\]

An approximate solution in the case of the AllInAs system is obtained by recognizing that:

$$\frac{\Phi_s}{q} + an_i = \frac{\Delta E_s}{q}$$  \[2.4\]

This estimate is good for AllInAs/GaInAs since $\Phi_s=0.6eV$, $\Delta E_s=0.52eV$, and $an_i=0.20eV$ for $n_i=1x10^{12}cm^{-2}$. The error in this approximation increases with $n_i$, but even at high channel concentrations, say $3x10^{12}cm^{-2}$, the approximation (equation [2.4]) is still acceptable. With the help of equation [2.4], equation [2.3] may be re-written (for the case when $V_s=0$) as:

$$E_2 \cdot d_2 = E_1 \cdot d_1$$  \[2.5\]

Substituting equation [2.2] into [2.5]:

$$E_2 \cdot d_2 = \frac{q n_i}{\varepsilon_2} \cdot d_1$$  \[2.6\]
Chapter 2

Next, a relationship between $E_2$ and $n_s$ is developed.

If the donor impurities are assumed to be fully ionized, then the field lines of the ionized charges in the donor layer are terminated by either charge in the channel or charge at the surface. From Gauss’ Law, the electric fields emanating from the planar-doping region can be described by:

$$E_2 + E_1 = \frac{qn_s}{\varepsilon_2}$$  \[2.7\]

which with the help of equation [2.2] may be written as:

$$E_2 = \frac{qn_s}{\varepsilon_2} - \frac{qm_1}{\varepsilon_2}$$  \[2.8\]

Next substitute equation [2.8] into [2.6] and solve for $n_s$, and one gets:

$$n_s = \frac{d_1}{d_1 + d_2} \cdot n_d$$  \[2.9a\]

or equivalently,

$$n_s = \frac{n_d}{1 + \frac{n_d}{d_1}}$$  \[2.9b\]

Equation [2.9] is the called the Lever Rule. Clearly, the barrier layer must be larger than the spacer layer in order to transfer the majority of the free carriers in the donor layer to the channel. In the limit as $\frac{n_d}{d_1}$ is made very large, one obtains:

$$\lim_{{\frac{n_d}{d_1} \rightarrow \infty}} n_s = 1$$  \[2.10\]

For a high barrier-to-spacer thickness ratio, the model predicts that all the free electrons in the donor layer will transfer to the channel.
Chapter 2

2.2 Access Region

There are two design criteria for the access regions of the JHEMT device. First, the access region sheet resistance must be minimized in order to minimize the source and drain parasitic transit delay. Second, the maximum current through the device should not be limited by the available charge in the access region.

Design Philosophy #A1: Minimize Source and Drain Transit Delays

The effects of the source and drain resistance on the total transit delay of the device may be seen by using nodal analysis on a simplified small-signal equivalent circuit model to obtain:

\[ \tau_{\text{total}} = \tau_{\text{int, intrinsic}} + \tau_{\text{parasitic}} \]  \hspace{1cm} [2.11]

\[ \tau_{\text{total}} = \frac{(C_{ps} + C_{gd})}{g_{ms}} + \frac{(C_{ps} + C_{gd}) \cdot (R_s + R_d)}{g_{ms} \cdot R_{de}} + C_{gd} \cdot (R_s + R_d) \]  \hspace{1cm} [2.12]

The first term is the intrinsic delay associated with charging the device capacitance. The second two terms correspond to the parasitic time delay through the drain and source resistance, and are particularly important when the gate length is reduced. The source resistance may be expressed as:

\[ R_s = R_{s,\text{access}} + R_c = R_{s,\text{hac}} \cdot \left( \frac{L_g}{W_s} \right) + R_c \]  \hspace{1cm} [2.13]

where the first term is the resistance associated with the access region and the second term is the contact transfer resistance which is addressed in Chapter 3. Clearly, a similar expression may be written for the drain resistance. To minimize the parasitic transit delay, the access
Chapter 2

Sheet resistance \( R_{sh,acc} \) must be minimized. The channel sheet resistance in the access region is given by:

\[
R_{sh,acc} = \frac{1}{(q \cdot \mu_e \cdot n_{s,acc})} \left( \frac{\Omega}{\text{square}} \right). \tag{2.14}
\]

Therefore, this design requirement demands both high electron mobility and high electron concentration for minimal sheet resistance in the access region.

Design Philosophy #A2: Maximize the Current Drive Capability

The higher gate barrier of the p+-AlInAs JHMET allows a higher forward turn-on voltage and, thus, the device may operate in enhancement mode. In such cases, as the gate is biased more positively, additional electrons are induced in the channel (under the gate) and the channel current increases, until one of the following occurs: 1) the gate diode turns on, 2) parasitic MESFET conduction occurs in the AlInAs donor layer, or 3) the electron velocity saturates in the access regions (limiting the available current through the device).

The first two points are addressed in section 2.3.3, but, the third point is the focus here. The current (per unit gate width) flowing from the source to the drain is continuous at any point along the channel and may be written as:

\[
I = q \cdot n_s(x) \cdot v(x) \quad \left( \frac{A}{cm} \right) \tag{2.15}
\]

where:
- \( q \) is the electron charge (C)
- \( n_s(x) \) is the sheet electron concentration \((cm^{-2})\)
- \( v(x) \) is the electron velocity \((cm/s)\).
Chapter 2

The maximum current which may flow through the access region is:

\[ I_{\text{max,acc}} = q \cdot n_{\text{acc}} \cdot v_{\text{sat}} \quad (A/cm) \]  \hspace{1cm} [2.16]

which is limited by the maximum number of free carriers \( n_{\text{acc}} \) in the region, all traveling at the electron saturation velocity \( v_{\text{sat}} \). Even if the gate is able to accumulate many more high velocity electrons, the access regions can not support the additional current. Consequently, the access region resistance rises dramatically\(^6\) causing device figures of merit (e.g. \( f_T, g_m \)) to degrade. Therefore, to suppress this phenomena, the access region must be designed for high electron concentration.

2.3 Gate Region

The design of the gate region begins with the derivation of a one-dimensional threshold voltage model which is used to predict threshold voltage. Next, a lumped-element model of the input impedance is discussed which leads to an expression for the gate resistance of the JHEMT. Finally, the design philosophies of the gate region are examined.

2.3.1 Threshold Voltage Model

In this section, a threshold voltage model is derived for the p\(^+\)-GaInAs/n-AllInAs/GaInAs JHEMT based on the linearized \( E_{\text{rx}}(n) \) function. Then a modified threshold voltage model is established in order to accurately predict the threshold voltage of the p\(^+\)-GaInAs/n-AllInAs/GaInAs JHEMTs discussed in Chapter 5.
Chapter 2

JHEMT Threshold Voltage Model

The planar-doped p\textsuperscript+\text/-GaInAs/n-AlInAs/GaInAs JHEMT is the structure chosen for this calculation. The energy band diagram and electric field of the structure are plotted in Figure 2.3. The following two assumptions are used in the following analysis: i) the donor impurities are fully ionized, and ii) the depletion approximation is valid.

Applying Gauss's Law to the channel, the electric field in the spacer layer may be related to the 2-DEG concentration as:

\[
E_1 = \frac{q n_1}{\varepsilon_1} \tag{2.17}
\]

where \( \varepsilon_1 \) is the dielectric constant in AlInAs. Also from Gauss's Law, the electric field in the barrier layer may related to both the donor sheet density and to the back depletion in the gate layer:

\[
E_2 = \frac{q n_A - q n_i}{\varepsilon_2} = \frac{q N_A x_p}{\varepsilon_2} \tag{2.18a}
\]

or more simply,

\[
N_A x_p = n_A - n_i \tag{2.18b}
\]

This expression is nothing other than a statement of charge conservation in a closed system with no emerging electric field. Equation [2.18b] may be solved to obtain the back depletion into the gate layer:

\[
x_p = \frac{n_A - n_i}{N_A} \tag{2.19}
\]

Next, by solving Poisson's equation under the gate contact, the potentials in the system are related by:
Figure 2.3. Energy band diagram and electric field profile of the planar-doped p+-GaInAs/n-AlInAs/GaInAs JHEMT. The donor layer comprises a sheet of donor atoms, \( n_d \) (cm\(^{-2}\)).
\[ \phi_{RC} - V_s - \frac{qN_A x_s^2}{2 \varepsilon_1} + \frac{\Delta E_F}{q} - \frac{q \kappa_1}{\varepsilon_2} + \frac{q \kappa_2}{\varepsilon_2} - \frac{\Delta E_F}{q} + E_{rx} = 0 \]  

[2.20]

where \( \varepsilon_1 \) is the dielectric constant in the GaInAs. Equation [2.20] may be simplified and rewritten by using equations [2.1] and [2.19], and recalling \( E_{ro} = 0 \) at 300K:

\[ \phi_{RC} - V_s - \frac{q(n_d - n_t)^2}{2 N_A \varepsilon_1} - \frac{q n_i d_2}{\varepsilon_2} + \frac{q m_t}{\varepsilon_2} (d_2 + d_1) + a n_t = 0 \]  

[2.21]

By definition, the threshold voltage is defined as the gate voltage at which the channel sheet concentration, \( n_t \), approaches zero, or:

\[ V_s = V_s \bigg|_{n_t \rightarrow 0} \]  

[2.22]

Applying [2.22] to [2.21] and solving for the threshold voltage, \( V_s \), one obtains:

\[ V_s = \phi_{RC} - \frac{q n_i^2}{2 N_A \varepsilon_1} - \frac{q n_i d_2}{\varepsilon_2} \]  

[2.23]

Equation [2.23] is the threshold voltage expression for the planar-doped p\(^+\)-GaInAs/n-AllInAs/GaInAs JHEMT. In this expression, \( \phi_{RC} \) is the energy from the conduction band to the Fermi level in the gate layer and is limited by the magnitude of the energy bandgap of the p-type semiconductor. For high acceptor concentrations, the second term in [2.23] vanishes and the threshold voltage varies linearly with the thickness of the barrier layer, \( d_2 \).

It is often desirable to replace the planar-doped donor layer by a thin, uniformly doped donor region. This allows high electron mobility for very thin spacer layers (see Chapter 5 and see reference\(^7\)). Instead of \( n_t \) (cm\(^{-2}\)) electrons in the donor layer, there are now
Chapter 2

\(N_s \cdot d_s \, (cm^{-2})\) electrons, where \(d_s\) is the finite thickness of the uniformly-doped donor layer. The new structure is called the pseudo-planar-doped (PPD) JHEMT, and the energy band diagram and electric field are plotted in Figure 2.4. The threshold voltage of the PPD-JHEMT easily derived and given by:

\[
V_a = \phi_{fc} - \frac{qN_c^2 d_s^2}{2N_s \epsilon_1} - \frac{qN_c d_s d_2}{\epsilon_2} - \frac{qN_c^2 d_s^2}{2\epsilon_2}
\]  

[2.24]

where the extra (fourth) term is attributed to the potential drop across the uniformly-doped donor region.

Clearly, for extremely high acceptor concentrations where the back depletion into the gate approaches zero \((x_s \to 0)\), one obtains the linear charge control model for Schottky HEMTs (where inherently the depletion in the metal is assumed to be zero). The linear control model first introduced by Delagebeaudeuf and Linh\(^6\) is

\[
qn_s = C_s \cdot (V_g - V_a)
\]  

[2.25]

where \(C_s\) is the 2-DEG capacitance per unit area, \(V_g\) is the applied gate voltage. Despite the model's simplicity, the linear behavior has been observed in AlGaAs/GaAs HEMTs at 12K\(^9\).

Modified JHEMT Threshold Voltage Model

In the previous threshold voltage model of the planar-doped and pseudo-planar-doped JHEMT, the purpose was to recognize the parameters which heavily influence the threshold voltage. Those model are extremely useful for that reason. However, the linearized \(E_{rx}(n_s)\) function (i.e. equation [2.1]) used in those models inaccurately
Figure 2.4. Energy band diagram and electric field profile of the pseudoplanar-doped (PPD) $p^+\text{GaInAs}/n\text{-AlInAs/GaInAs}$ JHEMT. Instead of a sheet of donor atoms, the donor region consists of $N_d$ (cm$^{-3}$) over a thickness, $d_n$. 
Chapter 2

describes the movement of the Fermi Level as the 2-DEG in the channel vanishes\textsuperscript{10}. Therefore, the threshold voltages calculated from the model inaccurately predict the threshold voltage. The purpose here is to derive a relation which accurately describes the threshold voltage in the p\textsuperscript{+}-GaInAs/n-AlInAs/GaInAs pseudo-planar-doped JHEMT.

The analysis begins with recalling the definition of threshold voltage in equation [2.22] and constructing the energy band diagram with a flat-band condition in the channel. The energy band diagram and electric field profile are shown in Figure 2.5. The following assumptions are made in this new model: i) all donor atoms are ionized and their field lines are terminated on acceptor charges in the gate region, ii) at threshold, the channel behaves as if it were intrinsic and the Fermi level has moved near the center of the bandgap (i.e. $E_{F,th} = E_g/2$), and iii) the depletion approximation is valid.

Using Gauss's Law, we may write the following charge relation:

$$N_d \cdot d_a = N_A \cdot x_p$$  \hspace{1cm} [2.26]

Again, the back depletion into the gate layer may readily be found when the channel is depleted of free carriers:

$$x_p = \frac{N_d \cdot d_a}{N_A}$$  \hspace{1cm} [2.27]

Now, the potential in the system may be summed to obtain:

$$\phi_{FC} - V_{th} - \frac{qN_A x_p^2}{2\varepsilon_1} + \frac{\Delta E_S}{q} - \frac{qN_A x_p}{2\varepsilon_2} - \frac{qN_d d_a^2}{2\varepsilon_2} - \frac{\Delta E_S}{q} - E_{F,th} = 0$$  \hspace{1cm} [2.28]

which after canceling terms and substituting equation [2.27], may be
Figure 2.5. Energy band diagram and electric field profile of the pseudoplanar-doped (PPD) p⁺-GaInAs/n-AllnAs/GaInAs JHEMT at the threshold condition. Solving the electrostatics with the position of the Fermi Level near the middle of the energy gap leads to the accurate calculation of the threshold voltage.
simplified to get:

\[ \phi_{FC} - V_{th} - \frac{qN_d^2d_n^2}{2N_A\varepsilon_1} - \frac{qN_d^2d_n}{\varepsilon_2}d_2 - \frac{qN_d^2d_n^2}{2\varepsilon_2} - E_{F,th} = 0 \quad [2.29] \]

Solving for the threshold voltage, one gets

\[ V_{th} = \phi_{FC} - \frac{qN_d^2d_n^2}{2N_A\varepsilon_1} - \frac{qN_d^2d_n}{\varepsilon_2}d_2 - \frac{qN_d^2d_n^2}{2\varepsilon_2} - E_{F,th} \quad [2.30] \]

Equation [2.30] is the modified threshold voltage equation for the pseudo-planar-doped JHEMT. As an example, the threshold voltage of wafer #V1307C (reported in Chapter 5) is estimated using the original charge control model and also calculated using the modified (flat-band condition) model. The results are listed in the table in Figure 2.6. The modified model developed here more accurately predicts the threshold voltage. The difference between the two models is accounted for by the relative position of the Fermi Level when the threshold condition is satisfied. In this case, the difference in the two models is \( E_s/2 = 0.375 \text{eV} \).

<table>
<thead>
<tr>
<th>Wafer V1307C</th>
<th>V\text{th}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure</td>
<td>50 Devices</td>
</tr>
<tr>
<td>Model</td>
<td>-0.76 to -0.81V</td>
</tr>
<tr>
<td>Charge Control Threshold Voltage Model</td>
<td>( -0.54V )</td>
</tr>
<tr>
<td>Model</td>
<td>Flat Band Threshold Voltage Model</td>
</tr>
</tbody>
</table>

Figure 2.6. Threshold voltage model comparison with actual devices measure on wafer V1307C. The Flat-Band Model (derived from Figure 2.7) more accurately predicts the threshold voltage due to the corrected assumptions about the position of the Fermi Level.