Chapter 2

Introduction to the Modelling of Semiconductor Devices

A complete discussion of the behaviour of semiconductor devices would require a knowledge of solid state physics and quantum mechanics, but for our purposes an intuitive understanding of their operation will be sufficient. §2.1 presents a simplified view of the physics of semiconductor devices and introduces the phenomena that make semiconductors technologically interesting. A complete discussion of the physics of semiconductor devices can be found in [31] and [34].

The remainder of this chapter contains an overview of semiconductor device modelling. §2.2 and §2.3 discuss the model itself — the nonlinear equations which govern semiconductors are developed in §2.2, and §2.3 outlines the boundary conditions, and some alternative formulations of the problem. §2.4 describes the form of the parameters which occur in the basic equations. §2.5 discusses the numerical solution of the equations.

2.1 Basic Operation of Semiconductor Devices

Traditionally materials have been classified according to their electrical properties as either conductors or insulators: under normal conditions conductors conduct electricity easily while insulators resist the current flow. This classification has been supplemented with another category, semiconductors, which allow current to flow more easily than insulators do, but are still relatively poor conductors.

The electrical properties of a material can be explained in terms of the intrinsic
concentration of conduction electrons in the solid [34]. (Conduction electrons are those electrons which are in an electron band so far removed from the nucleus that they can move about freely; valence electrons are electrons which occupy energy bands nearer the nucleus and are responsible for chemical bondings.) A typical semiconductor like silicon will have $10^{10}$ conduction electrons per cm$^3$; for a metal there will be roughly $10^{22}$ per cm$^3$ while for an insulator there will be just a few thousand conduction electrons per cm$^3$.

Semiconductors are important because a slight increase in the energy level (caused by applying an electric field or raising the temperature) will convert large numbers of valence electrons to conduction electrons. Obviously this will create gaps in the valence band where the valence electrons previously were. These gaps, which are called holes, can be treated as positively charged carriers. The movement of the holes and electrons creates a current. A slight change in the energy level changes the character of the material, and the semiconductor becomes a much better conductor.

In thermal equilibrium the number of holes is equal to the number of conduction electrons (this equilibrium concentration is denoted $n_i$). The electrical properties of the device can be changed by a process known as doping. There are two possibilities: either acceptor atoms, which can accept electrons and thus create holes, or donor atoms which can produce excess conduction electrons, can be implanted into the material. The concentration of the implanted carriers can be several orders of magnitude larger than the intrinsic level so doping changes the conductivity significantly. It is the arrangement of the impurities (the doping profile) which determines the behaviour of the semiconductor device.

The simplest example of a practical doping profile is the $pn$-junction illustrated in Figure 2.1. The net doping leaves the left ($p$-side) region with an excess of acceptors while the right ($n$-side) is dominated by donor atoms. The interface between the $n$-side
and the p-side is known as a pn-junction.

If a forward bias is applied to the diode (i.e., the voltage at the p-side is made positive with respect to the voltage at the n-side) a large current results, even for a small bias. But for a reverse bias (with the voltage at the p-side less than at the n-side) the current flow is negligible; even for a large negative voltage only a small (leakage) current results. The pn-junction can be considered as a switch, or valve, which allows a current to flow in one direction only.

The pn-junction is the simplest building block of semiconductor technology, but its behaviour illustrates the importance of the doping profile. A discussion of the pn-junction and other more sophisticated devices can be found in several texts, including [23] and [34].
2.2 Derivation of the Semiconductor Device Equations

The behaviour of a semiconductor device can be described in terms of the electrostatic potential, the carrier concentrations, and the current flow. A detailed derivation of the governing equations can be found in [31]. The briefer discussion which follows is based on the development found in [24]. We begin with Maxwell’s equations:

\[
\begin{align*}
\text{curl} H &= J + \frac{\partial D}{\partial t} \quad (2.1) \\
\text{curl} E &= -\frac{\partial B}{\partial t} \quad (2.2) \\
\text{div} D &= \rho \quad (2.3) \\
\text{div} B &= 0 \quad (2.4)
\end{align*}
\]

where \( E \) and \( D \) are the electromagnetic and displacement vectors, \( H \) and \( B \) are the magnetic field and induction vectors, \( J \) is the conduction current density and \( \rho \) is the space charge density. The electric field and the displacement can be related by \( D = \epsilon E \) where \( \epsilon \), the permittivity of the medium, is generally a 3x3 matrix. For most common applications of semiconductor devices it is reasonable to assume that the medium is isotropic so that \( \epsilon \) can be treated as a scalar.

2.2.1 Poisson’s Equation

The third of Maxwell’s equations (2.3) can be rewritten as Poisson’s equation by relating the electric field vector \( E \) to the electrostatic potential \( \psi \). Equation (2.4) can be satisfied by introducing a vector field \( A \) so that \( B = \text{curl} A \). Substituting this relation into (2.2) gives

\[
\text{curl}(E + \frac{\partial A}{\partial t}) = 0. \quad (2.5)
\]

For a simply connected domain, this implies that \( E \) can be expressed as:

\[
E = -\frac{\partial A}{\partial t} - \nabla \psi. \quad (2.6)
\]
Combining this relation with (2.3) gives (for homogeneous $\epsilon$)

$$\nabla(\epsilon \nabla \psi) = -\rho,$$

(2.7)

which is the usual form of Poisson's equation.

The space charge density $\rho$ can be expressed in terms of the elementary charge $q$ times the sum of the positively charged hole concentration $p$, the negatively charged electron concentration $n$ and the doping profile $C$:

$$\rho = q(p - n + C).$$

(2.8)

The doping profile $C$ is a function of position only. While this substitution can be viewed as a mathematical substitution, it is physically reasonable only if some conditions on $n$ and $p$ hold [31]. Equation (2.7) becomes (for constant $\epsilon$)

$$\Delta \psi = \frac{q}{\epsilon}(n - p - C).$$

(2.9)

2.2.2 The Continuity Equations

The principle of conservation of charge leads to two continuity equations. Applying the divergence operator to the first of Maxwell's equations (2.1) gives

$$\text{div}(\text{curl} H) = \text{div}J + \frac{\partial \rho}{\partial t} = 0.$$

(2.10)

The conduction current density $J$ can be split into two components $J_n$ and $J_p$: $J = J_n + J_p$. $J_n$ is the current created by the conduction of electrons, while $J_p$ is the current caused by the holes. The doping profile $C$ is time invariant, so that (2.10) becomes

$$-\text{div}J_p - q \frac{\partial p}{\partial t} = \text{div}J_n - q \frac{\partial n}{\partial t}.$$

(2.11)

This relation can be rewritten as two continuity equations by formally introducing a function $R$, so that

$$\text{div}J_p + q \frac{\partial p}{\partial t} = -qR$$

(2.12)
\[
\text{div} J_n - q \frac{\partial n}{\partial t} = qR. \tag{2.13}
\]

No new information is gained by rewriting (2.11) as two equations, but this formulation leads to a simple, physical interpretation of the function \( R \). \( R \) can be understood as the difference between the rate at which electron-hole pairs combine and the rate at which they are generated. For this reason \( R \) is usually called the recombination-generation function. \( R > 0 \) means that net recombination is occurring, while negative \( R \) means that pairs are being generated. \( R \) can be modelled physically using the methods of statistical physics; commonly used models are mentioned below.

2.2.3 The Carrier Transport Equations

Expressing the current densities \( J_n \) and \( J_p \) in terms of the electric field and the carrier concentrations is a cumbersome task. Boltzmann’s transport equation and the methods of statistical physics are used to derive, after lengthy calculations, fairly simple results. A more intuitive argument is used here to develop the same relations; a rigorous development, which discusses the necessary approximations and analyses the errors involved in making these assumptions, is given in [31].

Current flow in semiconductor devices has two main sources: the diffusion of electrons and holes due to the variation of their concentrations, and the drift of the carriers due to the variation of the electric field. The electron and hole current flows are determined by the sum of the diffusion and drift terms, so that

\[
J_n = J_n^{\text{diff}} + J_n^{\text{drift}} \tag{2.14}
\]

and

\[
J_p = J_p^{\text{diff}} + J_p^{\text{drift}}, \tag{2.15}
\]

where \( J_n^{\text{diff}}, J_p^{\text{diff}} \) are the current densities due to diffusion, and \( J_n^{\text{drift}}, J_p^{\text{drift}} \) are due to the drift.
Chapter 2. Introduction to the Modelling of Semiconductor Devices

The carriers diffuse in the direction of the gradient of the concentration from regions of higher concentration to regions of lower concentration. The total flux is proportional to this gradient and the charge per particle $q$, so

$$J_n^{diff} = qD_n \nabla n$$

and

$$J_p^{diff} = -qD_p \nabla p.$$  \hspace{1cm} (2.16)

(The signs of the right hand sides are chosen so that the diffusion coefficients $D_n$ and $D_p$ are positive.)

The drift current densities are simply the products of the charge per particle, the carrier concentrations, and the average drift velocities, denoted by $\nu_n^d$ and $\nu_p^d$ for electrons and holes respectively. Thus,

$$J_n^{drift} = qn\nu_n^d$$

and

$$J_p^{drift} = qp\nu_p^d.$$ \hspace{1cm} (2.18)

For moderate field strengths the drift velocities are proportional to the electric field, so that $\nu_n^d = \mu_n E$ and $\nu_p^d = \mu_p E$, where $\mu_p$ and $\mu_n$ are the hole and electron mobilities, respectively.

These relations for the diffusion and drift components of the current allow us to write:

$$J_n = qD_n \nabla n + q\mu_n nE$$

$$J_p = -qD_p \nabla p + q\mu_p pE.$$ \hspace{1cm} (2.20)

Finally, Einstein's relations $D_n = U_T \mu_n$ and $D_p = U_T \mu_p$ are used to relate the diffusion coefficients to the mobilities. Here $U_T = \frac{k_b T}{q}$ is the thermal voltage and $k_b$
is Boltzmann's constant. Einstein's relations are approximations with an acceptably small error if the semiconductor device is nondegenerate and the device temperature $T$ is spatially homogeneous [31]. These current relations can be written in terms of $\psi$ by assuming that the induction vector $B$ is independent of time, and substituting $E = -\nabla \psi$.

A more rigorous development produces the same results, but provides more insight into the physical assumptions which are necessary for (2.20) and (2.21) to hold. These equations neglect any current not caused by diffusion or the electric field. In particular, temperature effects and the variation of the intrinsic carrier concentration, which may be important for some applications, are neglected. Also, the superposition of terms in (2.20) and (2.21) is convenient but not necessarily correct.

2.3 The Semiconductor Equations

The basic semiconductor equations which will be used throughout the remainder of this thesis are collected in this section. The boundary conditions which supplement these equations are discussed in §2.3.1, and the equations are expressed in terms of nondimensional variables in §2.3.2. §2.3.3 presents two alternative formulations of the equations.

The major assumptions which are necessary for these basic equations to be valid are: that the current flow is caused only by diffusion and the drift caused by the electric field; that Einstein's relations hold; that the device temperature $T$ is constant; and that the induction vector $B$ is independent of time. These assumptions hold for the majority of applications. The following system of partial differential equations is obtained:

\begin{align*}
\epsilon \Delta \psi &= q(n - p - C) \\
\text{div} J_n - q \frac{\partial n}{\partial t} &= qR
\end{align*}
\[
\text{div} J_p + q \frac{\partial p}{\partial t} = -q R \\
J_n = qD_n \nabla n - q \mu_n n \nabla \psi \\
J_p = -qD_p \nabla p - q \mu_p p \nabla \psi.
\] (2.24, 2.25, 2.26)

Substituting (2.25, 2.26) into (2.23, 2.24) results in a system of three nonlinear partial differential equations for \( \psi, n \) and \( p \).

These equations were first presented by Van Roosbroeck [38] in virtually the same form in 1950. We will consider only the steady state semiconductor problem, obtained by setting the time derivatives to zero.

2.3.1 Boundary Conditions

Semiconductor devices are intrinsically three dimensional but the basic equations can also be posed in one or two dimensions. For a one dimensional problem the boundary of the device is trivially represented by just two points; for multi-dimensional problems the boundary \( \partial D \) of the domain \( D \) is considered to be piecewise smooth. \( \partial D \) can be split into two parts: \( \partial D = \partial D_p \cup \partial D_a \), where \( \partial D_p \) corresponds to the actual, physical device boundaries, and \( \partial D_a \) consists of those parts of the boundary which are artificially introduced to limit the size of the domain. The physical part of the boundary consists of insulating boundaries and contacts.

At insulating and artificial boundaries the components of the electric field and current densities normal to the boundary vanish:

\[
E \cdot r|_{\partial D_N} = J_n \cdot r|_{\partial D_N} = J_p \cdot r|_{\partial D_N} = 0
\] (2.27)

where \( r \) is the unit outward normal to \( \partial D \), and \( \partial D_N \) is the union of the insulating and artificial boundaries. Homogeneous Neumann conditions hold on \( \partial D_N \):

\[
\frac{\partial \psi}{\partial r}|_{\partial D_N} = \frac{\partial n}{\partial r}|_{\partial D_N} = \frac{\partial p}{\partial r}|_{\partial D_N} = 0.
\] (2.28)
At an Ohmic contact thermal equilibrium and a vanishing space charge are assumed to hold, so \( np = n_i^2 \) and \( \rho = n - p - C = 0 \). \( \psi \) is assumed to be the sum of the built-in potential which is produced by the doping and the applied potential \( V_0 \). These assumptions lead to the Dirichlet conditions

\[
\begin{align*}
    n|_{\partial D_o} &= \frac{1}{2}(C + \sqrt{C^2 + 4n_i^2}) \\
    p|_{\partial D_o} &= \frac{1}{2}(-C + \sqrt{C^2 + 4n_i^2}) \\
    \psi|_{\partial D_o} &= U_T \ln\left(\frac{C + \sqrt{C^2 + 4n_i^2}}{2n_i}\right) + V_0.
\end{align*}
\] (2.29) \quad (2.30) \quad (2.31)

Here \( \partial D_o \) is the union of all contacts in the boundary. In general \( V_0 \) is a function of time, but for the stationary problem it is treated as a constant.

Other, less common, boundary conditions arise when semiconductor-oxide interfaces and Schottky contacts are considered. A thorough discussion of the boundary conditions can be found in [31].

### 2.3.2 Singular Perturbation Scaling

The behaviour of the dependent variables becomes more manageable if a suitable scaling is used, and the governing equations are expressed in terms of dimensionless quantities. There are, of course, several ways to do this [30]. The scaling which best isolates the structure of the equations is the singular perturbation scaling [24]. With this scaling, all lengths are scaled by a characteristic device length \( l \), and the potentials are scaled by the thermal voltage \( U_T \). The doping profile \( C(x) \), and the carrier concentrations, are scaled by \( C_m \), the maximum magnitude of \( C(x) \). Finally the carrier mobilities \( \mu_n \) and \( \mu_p \) are scaled to have order of magnitude 1. Typical scaling values are shown in Table 2.1 [24].

The resulting scaled equations, with the scaled quantities denoted by the same
Chapter 2. Introduction to the Modelling of Semiconductor Devices

<table>
<thead>
<tr>
<th>Variable</th>
<th>Scaling Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$5 \times 10^{-3} cm$</td>
</tr>
<tr>
<td>$\psi$</td>
<td>0.0259V</td>
</tr>
<tr>
<td>$n$</td>
<td>$10^{17} cm^{-3}$</td>
</tr>
<tr>
<td>$p$</td>
<td>$10^{17} cm^{-3}$</td>
</tr>
<tr>
<td>$n_i$</td>
<td>$10^{17} cm^{-3}$</td>
</tr>
<tr>
<td>$\mu_n$</td>
<td>$1000 cm^2 V^{-1} s^{-1}$</td>
</tr>
<tr>
<td>$\mu_p$</td>
<td>$1000 cm^2 V^{-1} s^{-1}$</td>
</tr>
</tbody>
</table>

Table 2.1: Typical Scaling Values

symbols as the unscaled variables were, are:

\[
\lambda^2 \Delta \psi = (n - p - C) \tag{2.32}
\]

\[
\text{div} J_p + \frac{\partial p}{\partial t} = -R \tag{2.33}
\]

\[
\text{div} J_n - \frac{\partial n}{\partial t} = R \tag{2.34}
\]

\[
J_n = \mu_n (\nabla n - n \nabla \psi) \tag{2.35}
\]

\[
J_p = -\mu_p (\nabla p + p \nabla \psi) \tag{2.36}
\]

where

\[
\lambda = \sqrt{\frac{e U_T}{q C_m \tau^2}}
\]

is a very small value, typically of the order of magnitude $10^{-3}$ to $10^{-5}$. Another small parameter $\delta^2 = n_i/C_m$ arises when the boundary conditions are correspondingly scaled but does not appear in the differential equations.

The equations are singularly perturbed since a small parameter $\lambda^2$ multiplies the highest derivative in Poisson's equation. The solution components are expected to have layers across junctions of the doping profile, since where $\psi$ is slowly varying $n-p \sim C(x)$ (as $\lambda \to 0$) and $C(x)$ has discontinuities at the junctions (see below). A numerical solution of the equations should account for this behaviour: a fine discretization will
be needed near the layers if the solution is to be accurately approximated there, but a
course mesh will suffice elsewhere.

2.3.3 Transformations of Dependent Variables

For both analytic and computational work there are possible advantages to using sets
of dependent variables other than ψ, n and p [30]. One transformation in common use
replaces the unscaled carrier concentrations n and p by the quasi-Fermi levels φn and
φp as follows:

\begin{align}
  n &= e^{\psi - \phi_n} \\
  p &= e^{\phi_p - \psi}.
\end{align}

(In terms of the scaled variables, this transformation is: \( n = \delta^2 e^{\psi - \phi_n}, \ p = \delta^2 e^{\phi_p - \psi} \).)

The basic (unscaled) equations (2.22 – 2.26) are transformed to:

\begin{align}
  \epsilon \Delta \psi - q(e^{\psi - \phi_n} - e^{\phi_p - \psi} - C(x)) &= 0 \\
  \text{div}(\mu_n e^{\psi - \phi_n} \nabla \phi_n) + R &= 0 \\
  \text{div}(\mu_p e^{\phi_p - \psi} \nabla \phi_p) - R &= 0,
\end{align}

so the continuity equations are now in divergence form. There is a physical interpre-
tation of φn and φp, but for our purposes it is reasonable to treat the transformation
simply as a change of variables.

The second transformation to the Slotboom variables u and v reads:

\begin{align}
  n &= e^\psi u \\
  p &= e^{-\psi} v,
\end{align}

or, equivalently:

\begin{align}
  u &= e^{-\phi_n}
\end{align}
\[ v = e^{\phi_u}. \]  

(2.45)

(The corresponding transformation in terms of the scaled variables is: \( n = \delta^2 e^\psi u, \)
\( p = \delta^2 e^{-\psi} v. \)) The resulting (unscaled) equations are:

\[ \epsilon \Delta \psi - q(e^\psi u - e^{-\psi} v - C(x)) = 0 \]  

(2.46)

\[ \text{div}(\mu_n e^\psi \nabla u) - R = 0 \]  

(2.47)

\[ \text{div}(\mu_p e^{-\psi} \nabla v) - R = 0. \]  

(2.48)

Both transformations "regularize" the problem, in that \( \phi_n, \phi_p, u \) and \( v \) are all slow
variables (they vary slowly through the junctions), and the equations are no longer
singular, singularly perturbed. The Slotboom variables are attractive mathematically
because (2.47) and (2.48) are linear in \( u \) and \( v \), while (2.40) and (2.41) are quasilinear
in \( \phi_n \) and \( \phi_p \); further (2.47) and (2.48) are self adjoint for given values of \( \psi \) and \( R \). The
dynamic ranges of \( (\phi_n, \phi_p) \) and \( (n, p) \) are moderate, and successful numerical computations have been performed using both sets of variables [30]. The Slotboom variables
are poorly scaled; this formulation is generally only useful for analytic investigations.

The typical values which the scaled variables take on are summarized in Table 2.2

[30].

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>([10^{-20}, 10])</td>
</tr>
<tr>
<td>( p )</td>
<td>([10^{-20}, 10])</td>
</tr>
<tr>
<td>( \phi_n )</td>
<td>([0, 193.4])</td>
</tr>
<tr>
<td>( \phi_p )</td>
<td>([0, 193.4])</td>
</tr>
<tr>
<td>( u )</td>
<td>([1, 10^{64}])</td>
</tr>
<tr>
<td>( v )</td>
<td>([10^{-64}, 1])</td>
</tr>
</tbody>
</table>

Table 2.2: Range of Dependent Variables
2.4 Parameter Models

The form of the parameters arising in the development of the semiconductor equations needs to be specified before further analysis can be performed. The modelling of these physical parameters is a difficult task involving research in solid state physics [31]. Practically the parameters are known only as the (numerical) solution to another set of partial differential equations, or result from fitting measured data to theoretical models. The functions which are described below (in terms of the unscaled variables) are commonly used, but there is still some uncertainty in these models. A complete discussion of the parameter models is contained in [31].

2.4.1 The Doping Profile

The behaviour of a semiconductor device is mainly determined by its doping profile, the concentration of donor and acceptor impurities in the device. The creation of the doping profile is a complicated process involving ion implantation, diffusion and thermal oxidation.

The resulting doping profile $C(x)$ is the net difference of electrically active donors and acceptors. Regions in which $C(x) > 0$ (i.e., the concentration of donors exceeds the concentration of acceptors) are termed $n$-domains, while regions where $C(x) < 0$ are called $p$-domains. The boundary between an $n$-domain and a $p$-domain is called a $pn$-junction. The doping profile varies extremely rapidly near $pn$-junctions but is relatively constant away from the junctions. For mathematical convenience junctions are often modelled by abrupt transitions, so that $C(x)$ has jump discontinuities. The discontinuities are nonphysical but they provide a good approximation to many realistic profiles, and are convenient for mathematical analysis. The doping function can also be modelled using error functions and decaying exponentials.
\( C(x) \) typically takes on values in the range \((-1.0 \times 10^{18}, 1.0 \times 10^{18})\).

2.4.2 The Recombination-Generation Function

The derivation of the semiconductor device equations introduced the recombination -
generation function \( R \) in a purely formal way. This function has a physical interpreta-
tion as the net difference of the rates of recombination and generation of electron- hole
pairs. Recombination describes the neutralization of a hole when a conduction electron
becomes a valence electron, while generation refers to the opposite process in which a
valence electron becomes a conduction electron.

In thermal equilibrium the concentrations of electrons and holes will fluctuate but,
overall, there will be a dynamic balance so \( R \equiv 0 \). The application of an external
stimulus will disturb the equilibrium, and various processes occur which attempt to
create a new steady-state. If excess carriers have been created then recombination
processes prevail \( (R > 0) \) while generative processes dominate \( (R < 0) \) if carriers were
removed.

The recombination and generation is affected by several energy transition processes.
Each of these component processes can be analysed using techniques from quantum
mechanics to produce an overall model of \( R \) [31]. The simplest energy transition process,
the two particle transition, is modelled by the Schockley-Read-Hall term

\[
R_{SRH} = \frac{np - n_i^2}{\tau_p^l(n + n_i) + \tau_n^l(p + n_i)}
\]  

(2.49)

while the Auger recombination- generation term

\[
R_{AU} = (C_n^{AU}n + C_p^{AU}p)(np - n_i^2)
\]  

(2.50)

models the three particle transition. Here \( \tau_p^l \) and \( \tau_n^l \) depend on the doping profile and
represent the hole and electron lifetimes. Typical values for \( \tau_n^l \) and \( \tau_p^l \), as well as for
the constants \( C_n^{AU} \) and \( C_p^{AU} \) are given in Table 2.3.
Chapter 2. Introduction to the Modelling of Semiconductor Devices

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Typical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau_n )</td>
<td>( 10^{-6} \text{s} )</td>
</tr>
<tr>
<td>( \tau_p )</td>
<td>( 10^{-6} \text{s} )</td>
</tr>
<tr>
<td>( n_i )</td>
<td>( 10^{10} \text{cm}^{-3} )</td>
</tr>
<tr>
<td>( C_n^{AU} )</td>
<td>( 2.8 \times 10^{-31} \text{cm}^6\text{s}^{-1} )</td>
</tr>
<tr>
<td>( C_p^{AU} )</td>
<td>( 9.9 \times 10^{-32} \text{cm}^6\text{s}^{-1} )</td>
</tr>
</tbody>
</table>

Table 2.3: Values of parameters occuring in \( R \)

Other recombination and generation processes which are important only in special situations, or whose contribution is small, are omitted from this discussion but are described in [31]. Many models simply use \( R = R_{SRH} \). For some computer models, and for much analytic work, it is assumed that \( R \equiv 0 \). Strictly this is valid only for thermal equilibrium, but it is also reasonable for problems 'near' equilibrium.

2.4.3 The Carrier Mobilities

The carrier mobilities \( \mu_n \) and \( \mu_p \) occur as multiplicative factors in the expressions for the current components \( J_n \) and \( J_p \), so accurate modelling of the mobilities is important. Unfortunately, the physical mechanisms involved are difficult to model exactly, so many commonly used models are based only on experimental data [24]. The carrier mobilities are related to the electron and hole relaxation times \( \tau_n^r \) and \( \tau_p^r \) by

\[
\mu_n = \frac{q \tau_n^r}{m_n^*} \quad (2.51)
\]

\[
\mu_p = \frac{q \tau_p^r}{m_p^*} \quad (2.52)
\]

where \( m_n^* \) and \( m_p^* \) are the effective electron and hole masses respectively. The relaxation times describe the average time between scattering events for the carriers. The carriers are scattered by thermal lattice vibrations, ionized and neutral impurities, surfaces and other carriers. Models for each effect are described in [31]; the combined effect can be
found by an appropriate averaging of the individual events. The carrier mobilities are further reduced by the saturation of the drift velocity due to lattice vibration. The electron mobility generally varies between 50 $cm^2 V^{-1} s^{-1}$ and 1500 $cm^2 V^{-1} s^{-1}$, while the hole mobility $\mu_p$ ranges between 50 $cm^2 V^{-1} s^{-1}$ and 500 $cm^2 V^{-1} s^{-1}$.

2.5 The Numerical Solution of the Semiconductor Equations

The previous sections have developed and discussed a mathematical model of semiconductor devices. The model consists of a system of nonlinear partial differential equations which cannot be solved analytically, in general. To generate solutions for arbitrary conditions numerical techniques must be used.

This section addresses the two major issues arising from the numerical problem. First, a stable discretization of the basic equations is derived using a finite difference scheme. The techniques which are commonly used to solve the nonlinear equations are outlined in §2.5.2. We consider the basic equations in their scaled form (2.32 – 2.36) throughout the remainder of this thesis.

2.5.1 Discretizing the Equations

There are several classical methods which can be used to discretize the semiconductor equations [6]. Finite elements, finite differences and the finite box method (a generalization of the finite difference method) have all been successfully used to generate discretizations. The most commonly used approach is a specialized finite difference scheme known as the Scharfetter-Gummel discretization [32]. This approach will be detailed below for the one dimensional case.

We consider a set of mesh points $\{x_i\}_{i=0,1,N}$ with $x_0 = 0$ and $x_N = 1$ the boundary points. Again recall that the singularly perturbed nature of the equations dictates that
the mesh be highly nonuniform: away from the doping profile junctions the solution is smoothly varying and so a coarse grid is sufficient, but the solution changes rapidly through the junction layers so a fine grid is required in these regions. However, for an efficient solution (especially in the multigrid context) it is important to have a stable discretization even when the grid is coarse at the junctions.

The discretization of Poisson's equation with a small parameter is a standard problem which is treated in many publications, for instance [29]. We use the standard three point scheme for Poisson's equation on a nonuniform grid to give

\[
\frac{2\lambda^2}{h_i(h_{i-1} + h_i)}(\psi_{i+1} - \psi_i) - \frac{2\lambda^2}{h_{i-1}(h_{i-1} + h_i)}(\psi_i - \psi_{i-1}) = n_i - p_i - C(x_i)
\] (2.53)

where \( h_i = x_{i+1} - x_i \). The error involved in making the discretization is linearly proportional (with a very small proportionality constant) to the mesh spacing and the third partial derivatives of the potential. Note that as \( \lambda \to 0 \) the reduced solution away from the junctions is reproduced.

The treatment of the continuity equations is less standard. Although the small parameter \( \lambda^2 \) appears explicitly only in Poisson's equation, the entire system is singularly perturbed. (It is, in fact, singular, singularly perturbed, i.e. the reduced system obtained by setting \( \lambda = 0 \) cannot be solved directly for the reduced potential \( \psi \) since \( \psi \) does not appear in the algebraic equation \( 0 = n - p - C(x) \); see, for example [3].) Using standard symmetric differences produces a scheme which is not sufficiently numerically stable in the sense that \( \nabla \psi \) is large near junctions, making the off-diagonal terms in the symmetric discretization large. The transformation to the Slotboom variables \( (u, v) \) is regularizing, but these variables are poorly scaled and hence numerically unusable. However, it is useful for generating a discretization.

In one dimension

\[
J_n = \delta^2 \mu_n e^\psi \frac{du}{dx}
\] (2.54)
and discretizing the current conservation relation using a symmetric scheme we have

\[
\frac{J_{n,i+1/2} - J_{n,i-1/2}}{1/2(h_i + h_{i-1})} = R(x_i). \tag{2.55}
\]

Assuming that \( \mu_n, J_n \) and \( \frac{du}{dx} \) are constant on the interval \([x_i, x_{i+1}]\), (2.54) gives

\[
\frac{du}{dx} = \frac{J_n e^{-\psi}}{\delta^2 \mu_n},
\]

which can be integrated exactly:

\[
u_{i+1} - u_i = \int_{x_i}^{x_{i+1}} \delta^2 \mu_n^{-1} J_n e^{-\psi} dx
\]

\[
= -\frac{1}{\delta^2 \mu_n} \frac{J_{n,i+1/2}}{\delta \mu_n (\frac{du}{dx})_{i+1/2}} [e^{-\psi}]_{x_{i+1}}^{x_i}
\]

\[
= -\frac{1}{\delta^2 \mu_n} \frac{J_{n,i+1/2} h_i}{(\psi_{i+1} - \psi_i)} (e^{-\psi_{i+1}} - e^{-\psi_i}). \tag{2.57}
\]

This determines \( J_{n,i+1/2} \) in terms of \( n_{i+1}, n_i, \psi_{i+1} \) and \( \psi_i \). Substituting this into (2.55) we have:

\[
\frac{2\mu_n}{h_i(h_{i-1} + h_i)} (B(\psi_{i+1} - \psi_i)n_{i+1} - B(\psi_i - \psi_{i+1})n_i) - \\
\frac{2\mu_n}{h_{i-1}(h_{i-1} + h_i)} (B(\psi_i - \psi_{i-1})n_i - B(\psi_{i-1} - \psi_i))n_{i-1} = R(n_i, p_i), \tag{2.58}
\]

where \( B(x) = \frac{x}{e^x - 1} \).

A similar treatment of the continuity equation for holes gives

\[
\frac{2\mu_p}{h_i(h_{i-1} + h_i)} (B(\psi_i - \psi_{i+1})p_{i+1} - B(\psi_{i+1} - \psi_i)p_i) - \\
\frac{2\mu_p}{h_{i-1}(h_{i-1} + h_i)} (B(\psi_i - \psi_{i-1})p_i - B(\psi_{i-1} - \psi_i))p_{i-1} = R(n_i, p_i). \tag{2.59}
\]

The local truncation error for the discretized equations is proportional to the grid spacing. This commonly used discretization is known as the Scharfetter-Gummel scheme [32], and is an instance of exponential fitting [4].
2.5.2 Solving Nonlinear Systems of Equations

The basic semiconductor equations are of the form

\[ F(u) = 0 \]  \hspace{1cm} (2.60)

where

\[ F(u) = \begin{pmatrix} F_1(u_1, u_2, u_3) \\ F_2(u_1, u_2, u_3) \\ F_3(u_1, u_2, u_3) \end{pmatrix} \]  \hspace{1cm} (2.61)

Here \((u_1, u_2, u_3)\) represent the dependent variables \((\psi, n, p)\), \((\psi, \phi_n, \phi_p)\) or \((\psi, u, v)\). The solution of nonlinear systems of equations is a standard problem, discussed in detail in [28], for example. The two approaches which are commonly used for the semiconductor equations are Newton’s method and Gummel’s iteration.

Newton’s method is a classical iterative technique for solving nonlinear systems of equations. It generates a sequence of approximate solutions \(\{u^k\}\) starting from an initial point \(u^0\) in the following manner:

\[ u^{k+1} := u^k + du \]  \hspace{1cm} (2.62)

where \(du\) satisfies the equations

\[ J(u^k)du = -F(u^k) \]  \hspace{1cm} (2.63)

with \(J(u)\) being the Jacobian matrix of (2.60). Equation (2.62) determines a Newton correction direction \(du\) and (2.63) takes a unit step in this direction to define the new iterate \(u^{k+1}\).

Newton’s method is quadratically convergent near the solution of the nonlinear equations. However, it is guaranteed to converge only locally so that a good starting point is necessary to ensure convergence. Further, each iteration can be expensive since
it involves the evaluation of the Jacobian matrix, and the solution of a set of (possibly ill-conditioned) equations.

Modified Newton methods attempt to improve the convergence of the basic algorithm by changing the direction of the Newton correction and the length of the step taken in this direction. Quasi-Newton methods reduce the cost of each iteration by using approximations to the Jacobian matrix which are easily computable, and invertible. The application of these techniques to the semiconductor equations is detailed in [5].

One problem which arises when applying Newton's method to the semiconductor equations is that the Jacobian matrices can be ill-conditioned. To overcome this ill-conditioning the correction transformation technique has been suggested [30]. The correction transformation technique combines two sets of variables by using the quasi-Fermi variables \( \phi_n \) and \( \phi_p \) for calculations while implicitly linearizing in terms of the carrier concentrations \( n \) and \( p \). A detailed analysis of the conditioning can be found in [3].

Newton's method treats the three nonlinear equations simultaneously, but if the terms which couple the equations are not large then these equations can be solved sequentially using a nonlinear Gauss-Seidel iteration. This iteration was first applied to the semiconductor equations by Gummel [17], and has become known as Gummel's method.

Gummel's method works as follows. Given \( (\psi^k, n^k, p^k) \) at iteration \( k \) the next iterate \( \psi^{k+1} \) is determined by solving Poisson's equation:

\[
\lambda^2 \Delta \psi^{k+1} = n^k - p^k + C(x).
\] (2.64)

Using \( \psi^{k+1} \) the values of \( (n^{k+1}, p^{k+1}) \) are found by solving the continuity equations:

\[
\text{div} J_n(\psi^{k+1}, n^{k+1}) - R(\psi^{k+1}, n^k, p^k) = 0
\] (2.65)
\[ \text{div} J_p(\psi^{k+1}, p^{k+1}) + R(\psi^{k+1}, n^k, p^k) = 0. \] (2.66)

These equations are linear because \( R \) is evaluated using the values from the previous iteration. In this original form, Gummel's method can be shown to diverge in some cases [24]. However, by linearizing \( R \) and evaluating \( R \) at the new iterates these problems can be avoided.

Gummel's method is simple, but has been very valuable in practice. The iteration will usually converge, even starting from a fairly poor initial value [31]. Its main disadvantage is that its convergence rate is usually no better than linear, and it becomes extremely slow when the equations are strongly linked, e.g., when there is a high applied potential, and the carrier concentrations are large.

Several modifications have been suggested to improve the speed of the Gummel iteration. Hwang et al [20] suggest a block iteration which is strongly related to Gummel's method, but which attempts to include more of the linkages between equations. Several authors have suggested using Gummel's method to generate a good initial estimate for Newton's method [1].