

JPL Publication 95-10

A Theoretical Analysis of Steady-State Photocurrents in Simple Silicon Diodes

L. Edmonds

March 1995

NASA

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

155 pgs Includes cover, ii - vi & 1 - 149

The research described in this publication was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

"Copyright (c) 1995, California Institute of Technology.
U.S. Government Sponsorship is acknowledged."

PREFACE

A theoretical analysis solves for the steady-state photocurrents produced by a given photogeneration rate function with negligible recombination in simple silicon diodes, consisting of a uniformly doped quasi-neutral region (called "substrate" below) between a p-n junction depletion region (DR) and an ohmic contact (electrode). Special attention is given to conditions that produce "funneling" (a term used by the single-event-effects community) under steady-state conditions. Funneling occurs when carriers are generated so fast that the DR becomes flooded and partially or completely collapses. Some or nearly all of the applied voltage plus built-in potential normally across the DR is now across the substrate. This substrate voltage drop affects substrate currents. The steady-state problem can provide some qualitative insights into the more difficult transient problem. Chapter 6 discusses some similarities between the steady-state and transient problems.

The DR boundary (DRB) is defined by an equation, but can be recognized from computer simulation results by plotting electron and hole densities, against a spatial coordinate, together on the same graph. Such a plot shows a reasonably well defined boundary that separates a space-charge region from a quasi-neutral region. With the DRB reasonably well defined, DR and substrate voltage drops are also reasonably well defined, and quantify the extent of DR collapse and the strength of funneling. A collapsed DR can also be recognized by a small width.

It was found that the substrate can divide into two subregions, with one controlling substrate resistance and the other characterized by ambipolar diffusion. It was also found that steady-state funneling is more difficult to induce in the p^+/n diode than in the n^+/p diode. The carrier density exceeding the doping density in the substrate and at the DRB is not a sufficient condition to collapse a DR. A simple necessary condition for a DR collapse (or funneling) is derived in terms of ambipolar diffusion currents and is a statement regarding the spatial distribution of carrier generation. The condition is satisfied if carrier generation is sufficiently close to the DR, but does not require generation inside of the DR. Quantitative predictions agree well with computer simulation results.

PREFACE (continued)

This is the first rigorous (albeit steady-state) analysis of funneling in three dimensions, and may help to dispel some myths. Every point in the device lies on some equipotential surface, but a common misconception is that one such surface, called a "funnel", is distinguished from the others by containing the region where there is a substrate electric field. In this picture, the electric field is in a region that extends a "funnel length" from the DRB into the substrate. In reality, the electric field is not confined to such a region and there is no unambiguous funnel. The region containing the strongest substrate electric field is typically adjacent to the electrode, where the carrier-density-modulated conductivity is smallest. This is seen under transient as well as steady-state conditions. The total substrate voltage drop measures the extent of DR collapse, but the distribution of this potential within the substrate merely responds to the carrier-density-modulated conductivity. Selection of a surface to be called a funnel is arbitrary, and the concept of a funnel was not found to be useful. Another common misconception is that funneling requires that carriers be generated inside the DR. In reality, carriers generated outside but close to the DR can also induce funneling. This is also seen under transient as well as steady-state conditions.

The level of rigor accounts for the length of this analysis. Readers that are not interested in mathematical theory should be able to understand Chapters 1 and 6 without reading the other chapters.

CONTENTS

1. INTRODUCTION.....	1
2. PRELIMINARY DISCUSSION AND GOVERNING EQUATIONS.....	9
3. SUBSTRATE ANALYSIS: A SPECIAL CASE.....	17
3.1 Introduction.....	17
3.2 Solution for P and U.....	17
3.3 Solution for the Currents.....	20
3.4 The Nominal Ambipolar Approximation.....	21
3.5 A Generalized Ambipolar Approximation.....	22
3.6 Low-Injection-Level Conditions.....	26
3.7 Summary of Results for the p-Type Substrate.....	28
3.8 Analogous Results for the n-Type Substrate.....	30
4. SUBSTRATE ANALYSIS: THE GENERAL CASE.....	33
4.1 Introduction.....	33
4.2 Expressing Currents in Terms of $I_{e,1}$	33
4.3 Expressing $I_{e,1}$ in Terms of Γ	34
4.4 An Approximation for P and the Currents.....	37
4.5 A Mathematical Theorem.....	41
4.6 A Special Family of Generation Rate Functions.....	42
4.7 A Numerical Integration.....	47
5. THE COMPLETE SOLUTION.....	51
5.1 Introduction.....	51
5.2 Notation.....	51
5.3 Equation Summary for the n^+/p Diode.....	54
5.4 Algorithm for Constructing the n^+/p Diode I-V Curve.....	60
5.5 Equation Summary for the p^+/n Diode.....	62
5.6 Algorithm for Constructing the p^+/n Diode I-V Curve.....	68
5.7 A Necessary Condition for Saturation.....	70
6. NUMERICAL EXAMPLES AND CONCLUSIONS.....	73
6.1 Introduction.....	73
6.2 The One-Dimensional n^+/p Diode.....	74
6.3 The One-Dimensional p^+/n Diode.....	82
6.4 A Simple Three-Dimensional Diode.....	89
6.5 Conclusions.....	101

CONTENTS (continued)

APPENDIX A: THE DR EQUATIONS.....	107
A1 Introduction.....	107
A2 The n^+/p Junction.....	107
A3 The p^+/n Junction.....	112
 APPENDIX B: THE SPECIAL FUNCTION H.....	 113
B1 Introduction.....	113
B2 Definition of H when $Z_1 \geq 0$, $Z_2 > 0$, $Z_1 \neq Z_2$, and $1 + Z_1 - Z_2 \neq 0$...	114
B3 Some Inequalities.....	118
B4 Bounds for Case 1: $0 < Z_2 < Z_1$	122
B5 Bounds for Case 2: $0 \leq Z_1 < Z_2 < Z_1 + 1$	124
B6 Bounds for Case 3: $1 \leq Z_1 + 1 < Z_2$	125
B7 Some Additional Bounds for X_2	127
B8 Definition of $H(Z, Z)$ and $H(Z, 0)$ when $Z \geq 0$	128
B9 Asymptotic Forms.....	129
B10 Definition of $H(Z_1, Z_2)$ when $Z_1 \leq 0$, $Z_2 \leq 0$, and $1 + Z_1 - Z_2 \neq 0$...	133
B11 A Numerical Algorithm.....	133
B12 The Function Subprogram.....	134
 APPENDIX C: THE SPECIAL FUNCTION F.....	 141
 REFERENCES.....	 149
 FIGURES	
1.1 Qualitative Sketch of an n^+/p Diode.....	2
5.1 Qualitative Sketch of Both Diode Types.....	53
6.1 Comparison of I-V Curves for the n^+/p Diode.....	76
6.2 Comparison of Electron Density Predictions.....	78
6.3 n^+/p I-V Curves with Different Generation Locations.....	80
6.4 n^+/p I-V Curves Produced by a Reduced Generation Rate.....	81
6.5 Comparison of I-V Curves for the p^+/n Diode.....	83
6.6 Comparison Between n^+/p and p^+/n I-V Curves.....	84
6.7 Comparison of Hole Density Predictions.....	86
6.8 p^+/n I-V Curves with Different Generation Locations.....	88
6.9 A Simple Three-Dimensional Geometry.....	91
6.10 Comparison of I-V Curves for the 3D n^+/p Diode.....	98
6.11 Same as Fig.6.10 but without Recombination.....	100
6.12 Electron Density Versus v	102

1. INTRODUCTION

This publication analyzes simple silicon diodes exposed to steady-state photon irradiation. Funneling (a term used by the single-event-effects community [1]) occurs when carriers are generated in sufficient quantity near a p-n junction depletion region (DR) that the DR becomes flooded and partially, or completely, collapses. Some or nearly all voltage (including the built-in potential) normally across the DR is now across a substrate or epi layer, resulting in an electric field that enhances charge collection. This can occur under steady-state as well as transient conditions. The two types of conditions have some common qualitative characteristics, and concepts derived for the simpler steady-state problem can add physical insight into the more difficult transient problem. Theoretical transient models that exist at this time are unconvincing, and the primary motivation for the present steady-state analysis is to obtain physical and mathematical guidance for a future transient analysis. Therefore special attention is given to the extremely high irradiation intensities needed to produce steady-state funneling, such as might be produced by a laser having a pulse width longer than the device relaxation time. The analysis is not limited to such high-intensity conditions, but these are the only conditions under which the conclusions derived here differ significantly from those derived from the classical theory. Even when classical theory is known to apply, the treatment of three-dimensional geometries presented here may be found to be useful.

As shown in Figure 1.1, the simple silicon diode considered consists of a uniformly doped substrate between a p-n metallurgical junction (MJ) and an ohmic contact (electrode). The DR boundary (DRB) separates a strong space-charge region (the DR) from a quasi-neutral region. The simpler term "substrate" will refer to the quasi-neutral region from now on. Steady-state photogeneration occurs in the DR and/or substrate, and the generation rate density is assumed to be a known function (called the generation rate function) of the spatial coordinates. The figure shows an n^+/p device, but results are also given for the p^+/n device. The high-resistance region (HRR), ambipolar region (AR), and boundary (ARB) shown in the figure are discussed later.

The nonlinear drift-diffusion equations are simplified by assuming constant mobilities in the substrate (although electric

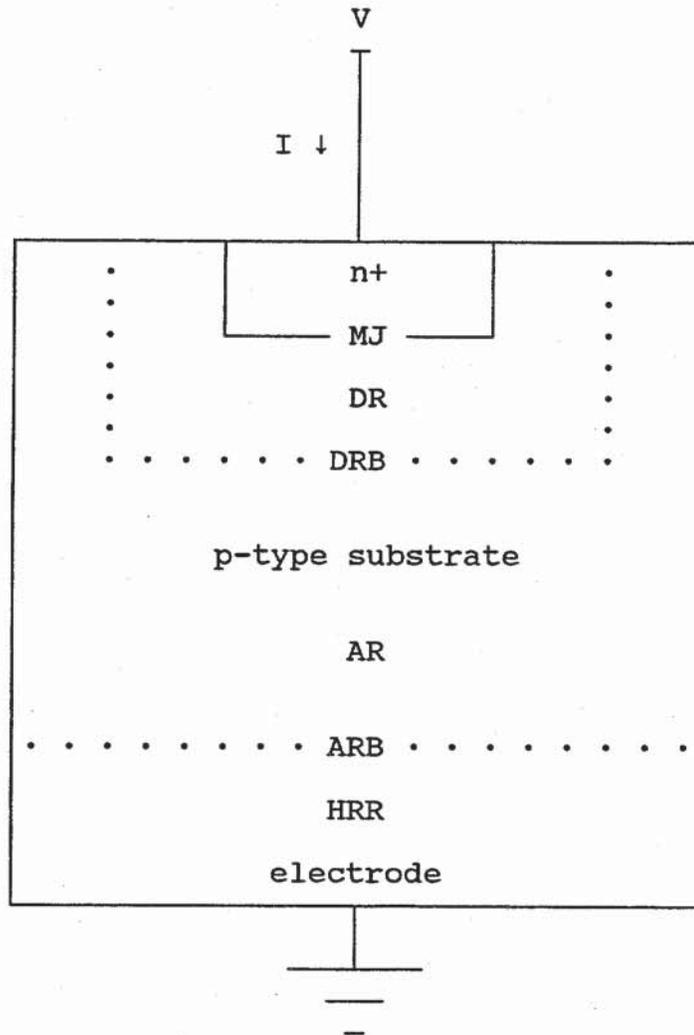


Figure 1.1: Qualitative sketch of an n+/p diode showing a metallogical junction (MJ), a depletion region (DR) and its boundary (DRB), an ambipolar region (AR) and its boundary (ARB), and a high-resistance region (HRR). The current I is positive when directed downward.

field dependent mobilities are used in the DR) and neglecting recombination (except at the electrode). From this point on, the analysis is fairly rigorous. Compared to the classical analysis, the analysis given here is more general in one sense but more limited in another. It is limited to cases where recombination can be neglected. It is more general in the sense that it applies to a wide range of operating conditions including (but not limited to) those that produce currents large enough for the classical law of the junction to break down, and that produce strong electric fields in the substrate. Furthermore, the analysis applies to arbitrary substrate geometries and does not require that the DRB be a single connected surface. It can be the union of any number of disconnected surfaces (i.e., an array of DRBs) providing that the same carrier density and potential boundary values are common to all surfaces. Similarly, the electrode can be the union of any number of disconnected ohmic contacts. However, if the DRB and/or electrode consist of several disconnected sections, currents through the individual sections are not solved. Sums of currents (summed over the individual sections) are solved.

The complete analysis consists of several distinct parts. One part, called the "DR analysis" solves the boundary value problem describing the DR. Another part, called the "substrate analysis" does the same thing for the substrate. The last part merely combines and solves the simultaneous equations provided by the other parts. Taken individually, the DR and substrate analysis are fairly general and can probably find applications in subjects other than an irradiated diode.

The DR analysis was originally presented in a publication that few people know about [2]. The results as originally presented were so complex that they were virtually unusable. These results are greatly simplified in Appendix A, and apply to a broad range of conditions (high or low injection levels, with or without velocity saturation). The substrate analysis (Chapters 3 and 4) applies to any substrate geometry and is more rigorous than analysis used in the past. It is never assumed in advance that one or another current component (electron or hole, drift or diffusion) can be neglected. It is sometimes concluded that one or another current component can be neglected, but the conclusion is derived (rather than assumed) and the conditions under which the conclusion is valid are quantified. Two special functions were found to be vital to the substrate analysis. These functions

are discussed extensively in Appendices B and C, which also contain subroutines for numerical evaluation.

Solutions are expressed in terms of equilibrium resistance (the resistance between electrode and DRB that would occur if there were no excess carriers), diffusion currents (predicted by the linear diffusion equation with simple boundary conditions), and a nameless quantity derived from the photogeneration rate function. These quantities implicitly contain the required geometric data and substitute for physical dimensions in the formal solutions (e.g., instead of specifying a length and area, we specify an equilibrium resistance). The advantage of this approach is that the equations are geometrically covariant, in the sense that the same equations are used for all geometries. Final numerical calculations are geometry specific and straightforward in one dimension. The three-dimensional case is made tractable by confining our attention to a special family of photogeneration rate functions, constructed so that all relevant functions of the spatial coordinates can be expressed as functions of a suitably chosen generalized coordinate (fitting is necessary if a given generation rate function does not belong to the family). Some manipulations then show how numerical estimates can be obtained from the same calculations that would be used in one dimension. The user must provide an equilibrium resistance estimate and a fitting function representing photogeneration. All other calculations, including diffusion current estimates, are first formally derived and then summarized in a "cookbook" recipe.

The equations used for the substrate are familiar to everyone, and an earlier publication [2] provides the complete list of DR equations. After listing these equations, the analysis is mathematical. This explains the scarcity of references. Although this is a mathematical analysis and very few physical arguments are used in the derivations, physical interpretations are given for some of the mathematical results. No apology is given for the fact that the analysis is lengthy. This is unavoidable because we are solving a set of simultaneous nonlinear partial differential equations in three dimensions. The final result will be a numerical algorithm for constructing the diode I-V curve corresponding to a given generation rate function. This can be done by a computer simulation, which can also treat diodes that are not simple, and does not require the user to provide the resistance estimate and fitting function discussed earlier. The value of this analysis is physical insight, including verification of the

statements made in the paragraphs below. Numerical examples in the last chapter provide a visual illustration of predicted physical results.

The present work finds that, when funneling is sufficiently strong, the ambipolar diffusion equation fails to provide a good approximation for the carrier density function, even when the predicted (via the ambipolar equation) carrier density is orders of magnitude greater than the doping density. The failure of this approximation is due to strong substrate electric fields. A more accurate equation is provided for quantitative calculations, but a simpler "generalized ambipolar approximation" is useful for visualization, and is described in the following way.

The substrate divides into two subregions (see Figure 1.1). Adjacent to the electrode is an HRR characterized by a small excess carrier density and strong electric field. This region forms because funneling-induced substrate fields drive minority carriers up from the electrode. There are virtually no replacement carriers supplied by the electrode, so the region is depleted of minority carriers. Quasi-neutrality insures that the region is also depleted of excess majority carriers. The conductivity is much less than in the high-density region above the HRR, so nearly all the substrate voltage drop is across the HRR. The region above the HRR is the AR and is characterized by a high carrier density and weak electric field. The ambipolar diffusion equation applies (approximately) to this region, but boundary conditions must be modified to account for the ARB that separates the AR from the HRR. It might be noted that the formation of an HRR and AR is very simple to derive in one dimension, if there is no photogeneration, and we assume in advance that the minority carrier current is negligible [3]. The present work derives this result in three dimensions, with photogeneration, and without the up-front assumption.

The HRR controls substrate resistance, while the ARB affects carrier density in the AR as if the electrode had been moved closer to the DRB. Furthermore, when funneling is sufficiently strong, the strong HRR electric field can drive nearly all minority carriers to the DRB. Replacing the electrode with a high-low junction, which blocks the minority carrier current, will have little effect because this current is blocked anyway. The device is in saturation during sufficiently strong steady-state funneling, i.e., nearly all liberated charge is collected. (This is one

distinction between the steady-state and transient cases. For the latter case, funneling is strong during part of the charge collection time at most, and the collected charge can be less than the total amount liberated.) It should be noted that even when funneling is not strong enough to produce saturation, it can still be important enough make the device I-V curve significantly different than the classical prediction.

The strong electric field in the HRR can affect mobility (velocity saturation) and it is reasonable to question the validity of ignoring this effect and assuming constant mobilities in the substrate. It turns out that the currents are insensitive to this effect, because the field in the HRR does not become this strong until the device is well into saturation. The carrier density in the HRR does respond to this effect and the minority carrier density changes from one negligible value to some other negligible value. The good quantitative agreement between predictions given here and those given by a computer simulation that includes electric field dependent mobilities, indicates that it is not necessary to use electric field dependent mobilities in the substrate.

A comparison is made between n^+/p and p^+/n diodes having the same geometry and doping (except that n-type and p-type are interchanged), subject to the same bias voltage (except for a change in polarity), and exposed to the same generation rate function. It was found that funneling is more difficult to induce in the n-type substrate. This observation goes beyond the simple fact that less mobile minority carriers are less responsive to a substrate electric field. In fact, the currents need not be greatly different and, depending on the bias voltage, either diode can have the larger current. The observation is that it is more difficult to create a substrate electric field in the p^+/n device. In one numerical example, the voltage across the p-type substrate was 1.63 volts, compared to only 0.11 volt across the n-type substrate. The DR was greatly collapsed in the former case, but nearly intact in the latter case, even though the carrier density greatly exceeded the doping density in the substrate and at the DRB (implying that this is not a sufficient condition for a DR collapse). A wide HRR occurred in the former case but not in the latter case; this compensated for the substrate voltage drops, so that the currents differed by less than 22%. A simple necessary condition for saturation (or a DR collapse) is derived in terms of ambipolar diffusion currents, and

is consistent with the conclusion that funneling is more difficult to induce in the p^+/n device.

Readers that are not interested in mathematical theory can go directly to Chapter 6 beginning on page 73.

2. PRELIMINARY DISCUSSION AND GOVERNING EQUATIONS

The analysis consists of several distinct steps. One step solves the equations describing the quasi-neutral region, which we have been calling the substrate. This is the region between the electrode (denoted S_1 for brevity) and the DRB (denoted S_2). In this context, "solve" means that the electron and hole currents are expressed in terms of the (unknown) carrier density and potential boundary values at S_2 . If we were treating a simple resistor, the equation $V=IR$ with R known would be called the solution. The solution for the semiconductor substrate is worked out in Chapters 3 and 4. Another step solves the equations describing the DR. Again "solve" means that currents are expressed in terms of boundary values or vice-versa. This step was already done in a previous publication. The results were very messy and are simplified in Appendix A. The third and last step combines and solves the simultaneous equations for the currents and boundary values. This step is analogous to using Kirchhoff's laws to solve the problem of two resistors in series, and is worked out in Chapter 5.

Because the DR analysis was already done, only the equations describing the quasi-neutral region need to be listed here. We start with the well-known equations which, under steady-state conditions with negligible recombination, reduce to

$$J_h = q D_h [- \text{grad } P - (P + p_0) \text{ grad } U/V_T] \quad (2.1a)$$

$$J_e = q D_e [\text{grad } N - (N + n_0) \text{ grad } U/V_T] \quad (2.1b)$$

$$\text{div } J_h = q g \quad (2.2a)$$

$$\text{div } J_e = - q g \quad (2.2b)$$

$$- \epsilon \text{ div grad } U = q (P - N) \quad (2.3)$$

$$D_e = V_T \mu_e, \quad D_h = V_T \mu_h \quad (2.4)$$

where

n_0, p_0 = equilibrium electron and hole densities, respectively
 N, P = excess electron and hole densities, respectively
 D_e, D_h = diffusion constants for electrons and holes, respectively
 μ_e, μ_h = mobilities for electrons and holes, respectively
 V_T = thermal voltage (about 0.026 volts at room temperature)
 q = elementary charge
 J_e, J_h = electron and hole current densities, respectively
 U = electric potential
 ϵ = dielectric constant
 g = generation rate function

The standard quasi-neutral approximation is obtained by regarding ϵ as sufficiently small compared to other relevant constants that the solutions to the equations can be approximated by the solutions obtained in the limiting case as ϵ approaches zero. In this limit, (2.3) is replaced with $P=N$ and (2.1) and (2.2) are used to solve for both P and U .

Boundary conditions should also be stated. The reference potential is chosen so that $U=0$ on S_1 , the semiconductor side of the electrode-semiconductor interface (contact potentials between electrodes and semiconductor will be included in Chapter 5), where we also have $P=0$. The values of P and U on S_2 are denoted P_2 and V_2 respectively, which are regarded as constants (in the spatial coordinates) on S_2 and represent some kind of spatial average on S_2 . All other boundary surfaces are insulated and assumed to be reflective for both electron and hole currents. This implies that the insulated boundaries are reflective for both P and U .

Although not essential, it is notationally convenient to be definite as to whether the substrate is an n- or p-type. Only one case need be considered in detail because analogous results apply to the other case. All discussions and analysis will refer to the p-type substrate. Final equations will be listed for the n-type case in Sections 3.8 and 5.5.

It is convenient to omit the equilibrium minority carrier

density n_0 in (2.1b). This term produces a theoretically predicted reverse current when the p-n junction is reverse-biased so that $P_2 \approx -n_0$. But this small current is not important because it is dominated by other currents (such as those associated with thermal generation/recombination in the DR) that are not included in this analysis. Therefore there is no compelling reason to keep the n_0 and we will leave it out.

The boundary value problem governing the p-type quasi-neutral region is now written as

$$J_h = q D_h [- \text{grad } P - (P + p_0) \text{ grad } U/V_T] \quad (2.5a)$$

$$J_e = q D_e [\text{grad } P - P \text{ grad } U/V_T] \quad (2.5b)$$

$$\begin{aligned} \text{div grad } P + \text{grad } P \cdot \text{grad } U/V_T + (P + p_0) \text{ div grad } U/V_T \\ = - g/D_h \end{aligned} \quad (2.6a)$$

$$\begin{aligned} \text{div grad } P - \text{grad } P \cdot \text{grad } U/V_T - P \text{ div grad } U/V_T \\ = - g/D_e \end{aligned} \quad (2.6b)$$

$$P = 0, \quad U = 0 \quad \text{on } S_1 \quad (2.7a)$$

$$P = P_2, \quad U = V_2 \quad \text{on } S_2 \quad (2.7b)$$

$$\text{grad } P \cdot \mathbf{n} = 0, \quad \text{grad } U \cdot \mathbf{n} = 0 \quad \text{on insulated boundaries} \quad (2.8)$$

where \mathbf{n} is the normal unit vector. The boundary value problem (2.5) through (2.8) is the mathematical definition of a "simple substrate" (for the p-type case under steady-state conditions). Although a simple substrate can only approximate a real physical system (at best), the equations themselves can be exactly solved for some special cases. An equation will be called exact if it is an exact mathematical result of these equations, regardless of

how well it represents a real physical system.

The objective of the next two chapters is to solve these equations so that the surface integrated currents are expressed in terms of P_2 and V_2 . Chapter 5 will solve for P_2 , V_2 , and all currents. The surface integrated currents are defined by

$$I_{h,i} \equiv \int_{S_i} \mathbf{J}_h \cdot d\mathbf{s} = -q D_h \int_{S_i} [\text{grad } P + (P + p_0) \text{ grad } U/V_T] \cdot d\mathbf{s} \quad (i=1,2) \quad (2.9a)$$

$$I_{e,i} \equiv \int_{S_i} \mathbf{J}_e \cdot d\mathbf{s} = q D_e \int_{S_i} [\text{grad } P - P \text{ grad } U/V_T] \cdot d\mathbf{s} \quad (i=1,2) \quad (2.9b)$$

$$I_{T,i} \equiv I_{h,i} + I_{e,i} \quad (i=1,2) \quad (2.9c)$$

where the unit normal vector in all surface integrals is an outer normal, i.e., directed away from the substrate interior. A surface integrated current is positive if positive charge moves toward the surface from the substrate interior.

Adding (2.5a) to (2.5b) and adding (2.6a) to (2.6b) produces a result that can be written as

$$\mathbf{J}_T = -\sigma \text{ grad } U_H \quad (2.10)$$

$$\text{div } \mathbf{J}_T = 0 \quad (2.11)$$

where

$$\mathbf{J}_T \equiv \mathbf{J}_e + \mathbf{J}_h \quad (2.12)$$

is the total current density and

$$\sigma \equiv q [\mu_e P + \mu_h (P + p_o)] = (q/V_T) (D_e + D_h) (P + A_o) \quad (2.13)$$

is the conductivity, with the constant A_o defined by

$$A_o \equiv D_h p_o / (D_e + D_h) . \quad (2.14)$$

The function U_H is defined by

$$U_H \equiv U - (2V_T/p_o) (p_o/2 - A_o) \ln(1 + P/A_o) . \quad (2.15)$$

Note that (2.10) and (2.11) are simply Ohm's law except that the "potential" is U_H instead of the actual potential U . The integrated form of Ohm's law is $V=IR$ or

$$V_2 - (2V_T/p_o) (p_o/2 - A_o) \ln(1 + P_2/A_o) = I_{T,1} R \quad (2.16)$$

where R is the resistance between S_1 and S_2 produced by the conductivity σ . This equation has limited computational applications because the carrier-density-modulated resistance R is unknown. The equation does have some applications, which will help to reach some conclusions in Sections 3.2 and 3.4.

Some constants and functions are defined below for later use. The equilibrium conductivity σ_o and ambipolar diffusion coefficient D^* are defined by

$$\sigma_o \equiv q \mu_h p_o = (q/V_T) D_h p_o \quad (2.17)$$

$$1/D^* \equiv (1/D_h + 1/D_e)/2 . \quad (2.18)$$

The unit function Ω_u and the function ϕ are defined by the boundary value problems

$$\text{div grad } \Omega_u = 0 \quad \text{in substrate} \quad (2.19a)$$

$$\Omega_u = 0 \quad \text{on } S_1 \quad (2.19b)$$

$$\Omega_u = 1 \quad \text{on } S_2 \quad (2.19c)$$

$$\text{grad } \Omega_u \cdot \mathbf{n} = 0 \quad \text{on insulated boundaries} \quad (2.19d)$$

$$\text{div grad } \phi = -g/D^* \quad \text{in substrate} \quad (2.20a)$$

$$\phi = 0 \quad \text{on } S_1 \quad (2.20b)$$

$$\phi = 0 \quad \text{on } S_2 \quad (2.20c)$$

$$\text{grad } \phi \cdot \mathbf{n} = 0 \quad \text{on insulated boundaries} \quad (2.20d)$$

Associated with these functions are the parameters R_0 , G_1 , and G_2 defined by

$$1/R_0 \equiv -\sigma_0 \int_{S_1} \text{grad } \Omega_u \cdot \mathbf{ds} = \sigma_0 \int_{S_2} \text{grad } \Omega_u \cdot \mathbf{ds} \quad (2.21)$$

$$G_i \equiv -q D^* \int_{S_i} \text{grad } \phi \cdot \mathbf{ds} \quad (i=1,2) \quad (2.22a)$$

The G's are related by

$$G_1 + G_2 = q \int_{\text{sub}} g d^3x . \quad (2.22b)$$

Each of these parameters has a physical interpretation. R_0 is the electrical resistance between S_1 and S_2 produced by the uniform equilibrium conductivity σ_0 . G_i ($i=1,2$) is the absolute value of the ambipolar diffusion current through S_i that would occur if the carrier density satisfied the ambipolar diffusion equation with S_1 and S_2 both acting as sinks for excess carriers (i.e., if $P=\phi$). These parameters are constants in the sense that they do not depend on spatial coordinates or on the boundary values P_2 or V_2 . However, they do depend on operating conditions. In addition to the obvious dependence that G_i has on g , there is also an implicit dependence due to the fact that the location of the boundary S_2 , which defines the geometry, can vary due to variations in the DR width. It will not be necessary to consider variations in the boundary S_2 until we get to Chapter 5. Chapters 3 and 4 will proceed as if the boundary location and boundary values are known and fixed. The parameters R_0 and the G 's are regarded as known when the boundary location is given. Chapter 4 will show how the G 's can be calculated from a particular type of function used to fit g .

3. SUBSTRATE ANALYSIS: A SPECIAL CASE

3.1 Introduction

We begin with a practice problem in which there is no photogeneration in the substrate. Although simpler than the more general case, this special case is far from trivial because carriers can be injected through S_2 . S_2 will be a p-n junction DRB in Chapter 5, but can presently be the boundary of any physical structure, because the boundary values P_2 and V_2 are arbitrary. In particular, it can represent a high-low junction, a forward biased p-n junction injecting minority carriers into the substrate, or a reverse-biased p-n junction injecting majority carriers into the substrate via photogeneration within the DR. Some concepts applicable to more general conditions are most easily discovered by starting with this problem, because the analysis is not burdened by a lot of mathematical complexity and an exact solution can be found. Of special interest is the formation of an HRR and AR (discussed later) when V_2 is large and positive (a p-type substrate is assumed). This situation (funneling) occurs if carriers are generated within a reverse-biased DR fast enough to flood it, causing it to collapse so that much of the applied plus built-in voltage is across the substrate.

The analysis to follow regards the location of S_2 and the boundary values P_2 and V_2 as given constants. The equilibrium resistance R_0 is regarded as known, so the currents are considered to be solved when expressed in terms of P_2 , V_2 , and R_0 .

3.2 Solution for P and U

By adding (2.6a) to (2.6b) while using $g=0$, we obtain

$$\text{div grad } [P + (P_0/2V_T) U] = 0 . \quad (3.1)$$

Comparing the boundary value problem satisfied by the expression in brackets to (2.19), we find that

$$P + (P_0/2V_T) U = \Omega \quad (3.2)$$

where

$$\Omega \equiv [P_2 + (p_0/2V_T) V_2] \Omega_u \quad (3.3)$$

is regarded as a known function of the spatial coordinates. Using (3.2) to eliminate U in either (2.6a) or (2.6b) gives

$$\text{div} [(P + p_0/2) \text{grad} (P - \Omega)] = 0 . \quad (3.4)$$

The solution to this equation is P satisfying

$$P + (p_0/2 - A) \ln(1 + P/A) = \Omega \quad (3.5)$$

where A is a constant. Substituting (3.5) into (3.4) verifies that (3.5) is a solution. The boundary conditions are satisfied at S_1 . The constant A is selected so that the boundary conditions are also satisfied at S_2 . Evaluating (3.5) at S_2 , we find that A satisfies

$$(p_0/2 - A) \ln(1 + P_2/A) = (p_0/2V_T) V_2 \quad (3.6)$$

and can be calculated from either

$$A = (p_0/2) [1 - (V_2/V_T) E] \quad \text{if } V_2 \neq -2V_T P_2/p_0 \quad (3.7a)$$

or

$$A = P_2 (e^{1/E} - 1)^{-1} \quad \text{if } V_2 \neq -2V_T P_2/p_0 \quad (3.7b)$$

where

$$E \equiv [\ln(1 + 2P_2/p_0)]^{-1} \quad \text{if } V_2 = 0 \quad \text{and} \quad P_2 > 0 \quad (3.8a)$$

$$E \equiv H(Z_1, Z_2) \quad \text{if } V_2 \neq 0 \quad \text{and} \quad V_2 \neq -2V_T P_2/p_0 \quad (3.8b)$$

$$Z_1 \equiv (V_T/V_2) (1 + 2P_2/p_0) , \quad Z_2 \equiv V_T/V_2 \quad (3.8c)$$

and the special function H is defined by

$$H(Z_1, Z_2) = E \quad \text{if and only if} \quad \exp(1/E) = (E-Z_1)/(E-Z_2) . \quad (3.9)$$

Equations (3.7a) and (3.7b) give the same result in theory, but (3.7b) should be used if $(V_2/V_T)E$ is so nearly equal to 1 that (3.7a) requires more numerical precision than is available. Otherwise, (3.7a) can be used.

Properties of the function H are discussed in Appendix B, which also contains a subroutine for numerical evaluation. Although not obvious from a casual inspection of (3.9), there is a problem if $1+Z_1-Z_2=0$. As $1+Z_1-Z_2$ approaches zero, $H(Z_1, Z_2)$ becomes positively or negatively infinite, depending on whether the approach is from above or below. This problem case occurs when $P_2+(p_0/2V_T)V_2=0$ so that $\Omega=0$. The solution given by (3.5) does not apply to this case and must be replaced with

$$(P + p_0/2)^2 = [(P_2 + p_0/2)^2 - (p_0/2)^2] \Omega_u + (p_0/2)^2 \quad \text{if } \Omega=0$$

which is easily verified by substituting it into (3.4). This problem case will occur if S_2 is an electrode ($P_2=0$) and shorted to S_1 ($V_2=0$). But even if there is photogeneration in the substrate, this case is still not very interesting because, according to (2.16), the terminal current is zero. Other than this uninteresting example, the problem case would be associated (at least in concept) with a forward-biased DR ($V_2 < 0$) with the forward biasing strong enough to produce a large voltage drop across the highly conductive substrate. The current would quickly destroy the device. The problem case is not expected in applications of interest, so we will always use the solution given by (3.5) with A solved from (3.7).

P is solved from (3.5) and U is solved from (3.2). The solution for P can be written more explicitly by defining another special function F by

$$F(X_1, X_2) = Y \quad \text{if and only if} \quad Y + (1 - X_1) \ln(1 + Y/X_1) = X_2. \quad (3.10)$$

Properties of F are discussed in Appendix C, which also contains a subroutine for numerical evaluation. Comparing (3.5) and (3.10), we get

$$P = (p_0/2) F(2A/p_0, 2\Omega/p_0) . \quad (3.11)$$

3.3 Solution for the Currents

By taking the gradients of (3.2) and (3.5) and combining equations we get

$$\text{grad } P = [(P+A)/(P + p_0/2)] \text{ grad } \Omega \quad (3.12)$$

$$\text{grad } U = (2V_T/p_0) [(p_0/2 - A)/(P + p_0/2)] \text{ grad } \Omega . \quad (3.13)$$

Substituting these gradients into (2.9) gives

$$I_{h,2} = - I_{h,1} = 2q D_h (1 - A/p_0) \int_{S_1} \text{grad } \Omega \cdot ds$$

$$I_{e,2} = - I_{e,1} = - 2q D_e (A/p_0) \int_{S_1} \text{grad } \Omega \cdot ds$$

and combining with (3.3) and (2.21) gives

$$I_{h,2} = - I_{h,1} = - (1 - A/p_0) (V_2 + 2V_T P_2/p_0)/R_0 \quad (3.14a)$$

$$I_{e,2} = - I_{e,1} = (D_e/D_h) (A/p_0) (V_2 + 2V_T P_2/p_0)/R_0 \quad (3.14b)$$

which, together with (3.7), completes the solution for the currents.

3.4 The Nominal Ambipolar Approximation

The behavior of P is easier to visualize if transcendental equation (3.5) is approximated by a simpler equation. The simplest approximation, which has some applications when $P_2 \gg p_0$, is the nominal ambipolar approximation obtained by neglecting U in (3.1) to get

$$P \approx P^* \quad (3.15)$$

where P^* is defined (when $g=0$) by the boundary value problem

$$\text{div grad } P^* = 0 \quad \text{in substrate} \quad (3.16a)$$

$$P^* = 0 \quad \text{on } S_1 \quad (3.16b)$$

$$P^* = P_2 \quad \text{on } S_2 \quad (3.16c)$$

$$\text{grad } P^* \cdot \mathbf{n} = 0 \quad \text{on insulated boundaries} \quad (3.16d)$$

Comparing (3.16) and (2.19), we find that

$$P^* = P_2 \Omega_u . \quad (3.17)$$

We can use (3.2) and (3.3) to conclude that the nominal ambipolar approximation (3.15) is valid if the ambipolar condition

$$P_2 \gg (p_0/2V_T) |V_2| \quad (\text{ambipolar condition}) \quad (3.18)$$

is satisfied.

Some of the older literature gives a misleading impression regarding ambipolar diffusion. The impression given is that electrons and holes interact so strongly, through their mutual attraction, that they move together and do not respond to applied fields. This picture accounts for U being absent in the equation governing P , but also predicts that $J_T=0$ (because electrons and holes move together). The assertion $J_T=0$ has also been supported by analysis of a strongly symmetric problem (cylindrical symmetry with no longitudinal flow). But such strong symmetry has some properties (e.g., the divergence of a bounded vector field uniquely determines the vector field) that do not apply to more general cases. The conclusion does not apply if the symmetry is weaker (e.g., cylindrical symmetry but with longitudinal flow) or if there is no symmetry. In the more general case, electrons and holes can move very differently from each other while maintaining quasi-neutrality, if carriers moving out of a volume element are replaced by others moving in. While it is true that the carrier density function is insensitive to weak applied fields, carrier motion is very responsive. This response can be seen from (2.16). R is insensitive to V_2 , so the total current is nearly linear in V_2 . Even when the ambipolar approximation is known to apply, we should avoid additional approximations derived from the idea that electrons and holes move together and independently of applied fields.

3.5 A Generalized Ambipolar Approximation

It is possible to modify the nominal ambipolar approximation to include some cases violating the ambipolar condition (3.18). We do assume throughout this discussion that $P_2 \gg p_0/2$. There are

four cases that can be considered. For the first case, V_2 is positive but small, where "small" means several times V_T . For the second case, V_2 is negative but small in absolute value. The nominal ambipolar approximation should apply to both of these cases. For the third case, V_2 is negative but large ($\gg V_T$) in absolute value. This case is not of practical interest. A physical arrangement producing this case is one in which S_2 represents a forward-biased p-n junction with an applied voltage strong enough to produce a large ohmic voltage drop across the highly conductive substrate. The large currents will quickly destroy the device. For the fourth case, V_2 is positive and large. This case can occur without destroying the device because a current limiting HRR forms (discussed below). A number of physical arrangements can produce the fourth case. Of special interest here is the one in which S_2 represents a reverse-biased p-n junction with photogeneration within the DR strong enough to collapse it, so that much of the applied plus built-in voltage is across the substrate (funneling). Given that $P_2 \gg p_0/2$, the fourth case is the only case of practical interest where the nominal ambipolar approximation fails. The objective of this section is to generalize the ambipolar approximation to include this case. The remainder of this section assumes that V_2 is positive.

An approximation for P can be derived by taking the gradient of (3.5) to get

$$\text{grad } P = [(P + A)/(P + p_0/2)] \text{ grad } \Omega . \quad (3.19)$$

It can be shown that a positive V_2 implies that A satisfying (3.6) also satisfies

$$0 < A < p_0/2 \quad \text{if } V_2 > 0 . \quad (3.20)$$

By assumption, $P_2 \gg p_0/2$. Therefore there is some region adjacent to S_2 where $P \gg p_0/2$ and $P \gg A$, so that the bracket in (3.19) is nearly unity, i.e., $\text{grad } P \approx \text{grad } \Omega$, implying that P and Ω differ (in this region) by an additive constant. The additive constant can be evaluated by noting that the region includes S_2 . The result is

$$P \approx \Omega - (p_0/2V_T) V_2 = [P_2 + (p_0/2V_T) V_2] \Omega_u - (p_0/2V_T) V_2 .$$

This equation is valid in a region sufficiently close to S_2 to satisfy $P \gg p_0/2$. Any points where the right side of the above equation is zero cannot be in this region. The ARB is mathematically defined to be the set of points where the right side of the above equation is zero, i.e., the constant Ω_u surface characterized by

$$\Omega_u = (p_0/2V_T) V_2 / [P_2 + (p_0/2V_T) V_2] \quad \text{defines ARB} . \quad (3.21)$$

The AR is mathematically defined to be the region between the ARB and S_2 . Excluding a transitional region adjacent to the ARB, the AR is characterized by $P \gg p_0/2$ so that

$$P \approx [P_2 + (p_0/2V_T) V_2] \Omega_u - (p_0/2V_T) V_2 \quad \text{in AR} . \quad (3.22a)$$

The HRR is mathematically defined to be the region between the ARB and the electrode S_1 . It can be shown from the exact equations that, excluding a transitional region adjacent to the ARB (where P can be several times $p_0/2$), the HRR is characterized by $P \ll p_0$ so that

$$P \approx 0 \quad \text{in HRR} . \quad (3.22b)$$

The HRR is characterized by a low conductivity ($\approx \sigma_0$, which is small compared to the conductivity in the AR) and a large (nearly all of V_2) potential drop when $V_2 \gg V_T$. This motivated the name "high-resistance region". This region limits the current so that a large V_2 can occur without destroying the device. The AR region is characterized by a high conductivity and small (several times V_T) potential drop. These are the conditions appropriate for ambipolar diffusion and motivated the name "ambipolar region".

We temporarily drop the assumption that V_2 is positive and define the generalized ambipolar approximation to be (3.22) when V_2 is positive and (3.15) otherwise. Reinstating the assumption

that V_2 is positive (so that the ARB exists), it is evident from (3.21) that the ARB becomes S_1 in the limit of small V_2 . In this same limit, the generalized approximation (3.22) reduces to the nominal approximation (3.15).

There is a physical explanation for the absence of excess carriers in the HRR. An HRR with sufficient width to be depleted of excess carriers (i.e., the HRR is distinguishable from the transitional region) forms when V_2 is large enough for the generalized ambipolar approximation to significantly differ from the nominal ambipolar approximation. But electric fields strong enough to make the nominal approximation fail are also strong enough to push electrons away from the electrode. The electrode supplies virtually no electrons, so there is a region near the electrode that is virtually depleted of electrons. Quasi-neutrality implies that this region is also virtually depleted of excess holes.

An alternate definition for the ARB, mathematically equivalent to (3.21), can be stated in terms of the slope of P . This alternate definition makes the ARB easier to visualize. The generalized and nominal ambipolar approximations predict the slope of P near S_2 to be given by

$$\text{grad } P \approx [P_2 + (p_0/2V_T) V_2] \text{ grad } \Omega_u \quad (\text{generalized}) \quad (3.23a)$$

$$\text{grad } P \approx P_2 \text{ grad } \Omega_u \quad (\text{nominal}) \quad (3.23b)$$

so that the generalized approximation predicts a steeper slope than the nominal approximation. The nominal approximation can be modified to give the generalized approximation by moving the sink boundary from the electrode to the ARB. Moving the sink boundary closer to S_2 produces a steeper slope. The ARB can be visualized (and defined) as the location where the sink must be placed to produce the correct (steeper) slope.

The generalized ambipolar approximation must be used with caution and should not be used in calculations that subtract nearly equal quantities and require high accuracy. For example, U is solved from (3.2) after P has been solved, but the exact solution must be used. Using the approximation for P will pre-

dict a zero electric field in the AR. This is not a good estimate of the electric field. The electric field is small in the AR only because the conductivity is correspondingly large, so even a small electric field is important and cannot be neglected. The generalized ambipolar approximation is an approximation for (3.2) and might be used instead of (3.2) (requiring that U be solved some other way), but cannot be used with (3.2). The approximation is useful for visualization, for predicting distinct regions where P has different behaviors, and for defining the ARB which separates these regions. But (3.11) is recommended for numerical calculations.

The final observation made here concerns the electron current. The conditions (large V_2) that result in the generalized ambipolar approximation being significantly different than the nominal approximation also result in A being extremely small. The electron current given by (3.14b) is extremely small. The physical explanation is the same as that given for the absence of excess carriers in the HRR. An electric field strong enough to cause the nominal approximation to fail is also strong enough to prevent electrons from reaching the electrode, so $I_{e,1} \approx 0$. This physical explanation also applies to the $g \neq 0$ case considered in Chapter 4. It is interesting to note that under large V_2 conditions, it makes no difference whether S_1 is an electrode or a high-low junction that blocks the electron current because $I_{e,1}$ is virtually zero anyway.

3.6 Low-Injection-Level Conditions

Low-injection-level conditions (LILC) occur when $P \ll p_0$ throughout the substrate. It is commonly assumed that LILC implies that the minority carrier diffusion equation (MCDE) gives a good approximation for P . It is interesting to determine whether this assumption is valid. It turns out that the assumption is invalid, but can still be used for the purpose of estimating total current. The meaning of this statement is explained below. It is also shown that the MCDE applies if and only if $A \gg P$.

Given LILC, a necessary condition for the MCDE to apply can be determined by comparing the MCDE-predicted gradients of P at S_1 and S_2 to the actual gradients. The solution to the MCDE, for steady-state conditions with negligible recombination/generation,

is $P_2\Omega_u$ (the same as the nominal ambipolar approximation). The predicted gradient of P at either boundary is $P_2\text{grad}\Omega_u$. The actual gradient is given by (3.19). Using (3.3) gives

$$\text{grad } P = (2/p_0) A [P_2 + (p_0/2V_T) V_2] \text{grad } \Omega_u \quad \text{at } S_1$$

$$\begin{aligned} \text{grad } P &= [(P_2 + A)/(P_2 + p_0/2)] [P_2 + (p_0/2V_T) V_2] \text{grad } \Omega_u \\ &\approx (2/p_0) (P_2 + A) [P_2 + (p_0/2V_T) V_2] \text{grad } \Omega_u \quad \text{at } S_2. \end{aligned}$$

One necessary condition for both of the above gradients to approximately equal $P_2\text{grad}\Omega_u$ is $A \gg P_2$, so that the coefficients on the two right sides will be nearly equal to each other. Another necessary condition is

$$(2/p_0) A [P_2 + (p_0/2V_T) V_2] \approx P_2$$

or

$$2 P_2/p_0 + V_2/V_T \approx P_2/A .$$

But $P_2/p_0 \ll 1$ and $P_2/A \ll 1$, so $|V_2/V_T| \ll 1$. We conclude that LILC are not sufficient for the MCDE to apply. It is also required that $|V_2| \ll V_T$.

A different line of reasoning will conclude that, given LILC, we can pretend that the MCDE applies, even if it really does not, providing that our interest is in total current. Given LILC, the minority carrier drift current is negligible compared to the majority carrier drift current. If the diffusion currents are on the order of, or larger than, the majority carrier drift current, then the minority carrier drift current is negligible compared to all other currents, and the MCDE applies (implying that $|V_2| \ll V_T$, which is consistent with the statement that majority carrier drift is as small as diffusion). If the diffusion currents are much smaller than the majority carrier drift current, then the MCDE does not apply. But we can pretend that it does, because nearly all current is majority carrier drift and error in the calculated diffusion current does not matter. Note that the MCDE

implies that $A \gg P$. Therefore, when calculating total current with LILC given, we can assume that $A \gg P$, even though the assumption may be wrong. Conditions under which the assumption is wrong are also conditions under which error in the assumption does not matter.

It was shown above that the MCDE implies that $A \gg P$. It is interesting that the implication also goes in the other direction. Given that $A \gg P$, we can expand the logarithms in (3.5) and (3.6) so that the equations reduce to

$$P + (p_0/2 - A) (P/A) \approx \Omega = [P_2 + (p_0/2V_T) V_2] \Omega_u$$

$$(p_0/2 - A) (P_2/A) \approx (p_0/2V_T) V_2$$

and combining equations to eliminate A gives $P \approx P_2 \Omega_u$.

3.7 Summary of Results for the p-Type Substrate

The results are now summarized for the p-type substrate. The equilibrium conductivity σ_0 is $q\mu_h p_0$ or $(q/V_T) D_h p_0$ where the equilibrium hole density p_0 can be equated to the doping density. The equilibrium resistance R_0 is the electrical resistance between S_1 and S_2 produced by the equilibrium conductivity. The constant A is calculated from either

$$A = (p_0/2) [1 - (V_2/V_T) E] \quad \text{if } V_2 \neq -2V_T P_2/p_0$$

or

$$A = P_2 (e^{1/E} - 1)^{-1} \quad \text{if } V_2 \neq -2V_T P_2/p_0$$

where

$$E \equiv [\ln(1 + 2P_2/p_0)]^{-1} \quad \text{if } V_2 = 0 \quad \text{and} \quad P_2 > 0$$

$$E \equiv H(Z_1, Z_2) \quad \text{if } V_2 \neq 0 \quad \text{and} \quad V_2 \neq -2V_T P_2/P_0$$

$$Z_1 \equiv (V_T/V_2) (1 + 2P_2/P_0) , \quad Z_2 \equiv V_T/V_2$$

and the special function H is defined in Appendix B. The two equations for A give the same result in theory, but the second should be used if $(V_2/V_T)E$ is so nearly equal to 1 that the first requires more numerical precision than is available. Otherwise, the first can be used. The exceptional case where A and E are undefined is mathematically possible but should not be encountered in practical applications. The currents are calculated from

$$I_{h,2} = -I_{h,1} = - (1 - A/p_0) (V_2 + 2V_T P_2/P_0)/R_0$$

$$I_{e,2} = -I_{e,1} = (D_e/D_h) (A/p_0) (V_2 + 2V_T P_2/P_0)/R_0 .$$

The above equations complete the solution for the substrate in the case where there is no photogeneration in the quasi-neutral region. But it is interesting to also look at the function P . The exact solution is given by either

$$P + (p_0/2 - A) \ln(1 + P/A) = \Omega$$

or

$$P = (p_0/2) F(2A/p_0, 2\Omega/p_0)$$

where

$$\Omega \equiv [P_2 + (p_0/2V_T) V_2] \Omega_u$$

with the unit function Ω_u defined by (2.19) and the special function F discussed in Appendix C. Approximations are available for P . First assume that $P_2 \ll p_0/2$. Then either majority carrier drift is the dominant current, or $P \approx P_2 \Omega_u$ and $A \gg P$. Now assume

that $P_2 \gg p_0/2$. For cases of practical interest such that $V_2 < 0$, the approximation is

$$P \approx P_2 \Omega_u \quad \text{if } V_2 < 0 \quad \text{and} \quad P_2 \gg p_0/2 .$$

If $V_2 \geq 0$, an approximation is obtained by defining the ARB to be the constant Ω surface characterized by

$$\Omega = (p_0/2V_T) V_2 .$$

The AR is the region between the ARB and S_2 while the HRR is the region between the ARB and S_1 . The approximation is

$$P \approx \Omega - (p_0/2V_T) V_2 \quad \text{in AR if } V_2 \geq 0 \quad \text{and} \quad P_2 \gg p_0/2$$

$$P \approx 0 \quad \text{in HRR if } V_2 \geq 0 \quad \text{and} \quad P_2 \gg p_0/2 .$$

The approximation is useful for visualization, but the solution in terms of F is recommended for numerical calculations.

3.8 Analogous Results for the n-Type Substrate

The analogous results are summarized for the n-type substrate. The equilibrium conductivity σ_0 is $q\mu_e n_0$ or $(q/V_T) D_e n_0$ where the equilibrium electron density n_0 can be equated to the doping density. The equilibrium resistance R_0 is the electrical resistance between S_1 and S_2 produced by the equilibrium conductivity. The constant A is calculated from either

$$A = (n_0/2) [1 + (V_2/V_T) E] \quad \text{if } V_2 \neq 2V_T P_2/n_0$$

or

$$A = P_2 (e^{1/E} - 1)^{-1} \quad \text{if } V_2 \neq 2V_T P_2/n_0$$

where

$$E \equiv [\ln(1 + 2P_2/n_0)]^{-1} \quad \text{if } V_2 = 0 \quad \text{and} \quad P_2 > 0$$

$$E \equiv H(Z_1, Z_2) \quad \text{if } V_2 \neq 0 \quad \text{and} \quad V_2 \neq 2V_T P_2/n_0$$

$$Z_1 \equiv - (V_T/V_2) (1 + 2P_2/n_0) , \quad Z_2 \equiv - V_T/V_2$$

and the special function H is defined in Appendix B. The two equations for A give the same result in theory, but the second should be used if $(V_2/V_T)E$ is so nearly equal to -1 that the first requires more numerical precision than is available. Otherwise, the first can be used. The exceptional case where A and E are undefined is mathematically possible but should not be encountered in practical applications. The currents are calculated from

$$I_{h,2} = - I_{h,1} = (D_h/D_e) (A/n_0) \cdot (V_2 - 2V_T P_2/n_0)/R_0$$

$$I_{e,2} = - I_{e,1} = - (1 - A/n_0) (V_2 - 2V_T P_2/n_0)/R_0$$

The exact solution for P is given by either

$$P + (n_0/2 - A) \ln(1 + P/A) = \Omega$$

or

$$P = (n_0/2) F(2A/n_0, 2\Omega/n_0)$$

where

$$\Omega \equiv [P_2 - (n_0/2V_T) V_2] \Omega_u$$

with the unit function Ω_u defined by (2.19) and the special function F discussed in Appendix C. Approximations are available for P . First assume that $P_2 \ll n_0/2$. Then either majority carrier drift is the dominant current, or $P \approx P_2 \Omega_u$ and $A \gg P$. Now assume that $P_2 \gg n_0/2$. For cases of practical interest such that $V_2 > 0$, the approximation is

$$P \approx P_2 \Omega_u \quad \text{if } V_2 > 0 \quad \text{and} \quad P_2 \gg n_0/2 .$$

If $V_2 \leq 0$, an approximation is obtained by defining the ARB to be the constant Ω surface characterized by

$$\Omega = - (n_0/2V_T) V_2 .$$

The AR is the region between the ARB and S_2 while the HRR is the region between the ARB and S_1 . The approximation is

$$P \approx \Omega + (n_0/2V_T) V_2 \quad \text{in AR if } V_2 \leq 0 \quad \text{and} \quad P_2 \gg n_0/2$$

$$P \approx 0 \quad \text{in HRR if } V_2 \leq 0 \quad \text{and} \quad P_2 \gg n_0/2 .$$

The approximation is useful for visualization, but the solution in terms of F is recommended for numerical calculations.

4. SUBSTRATE ANALYSIS: THE GENERAL CASE

4.1 Introduction

We now consider the general case in which there is photogeneration in the substrate. Unlike the special case in Chapter 3, exact solutions are not available for the general case. An exact analysis is used in Section 4.2 to express all currents in terms of $I_{e,1}$ (a p-type substrate is assumed here). Another exact analysis in Section 4.3 expresses $I_{e,1}$ in terms of a new unknown function Γ , which will eventually be approximated. Function Γ is constructed in such a way that an estimate of $I_{e,1}$ is insensitive to error in Γ . Section 4.4 gives an approximation for P , which is first used to approximate Γ , then $I_{e,1}$, and then the other currents. A mathematical theorem in Section 4.5, a suitable restriction on g discussed in Section 4.6, and a numerical integration discussed in Section 4.7 make the approximations computationally manageable. Unlike Chapter 3, this chapter does not end with summary sections, because the final equations (including those for the n-type substrate) are summarized in Sections 5.3 and 5.5.

4.2 Expressing Currents in Terms of $I_{e,1}$

By adding (2.6a) and (2.6b) while using (2.18), we obtain

$$\text{div grad } [P + (p_0/2V_T) U] = -g/D^* . \quad (4.1)$$

Comparing the boundary value problem satisfied by the expression in brackets to (2.19) and (2.20), we find that

$$P + (p_0/2V_T) U = \Omega + \phi \quad (4.2)$$

where

$$\Omega \equiv [P_2 + (p_0/2V_T) V_2] \Omega_u . \quad (4.3)$$

The two divergence equations (2.2a) and (2.2b) allow S_2 currents to be related to S_1 currents according to

$$I_{h,2} = q \int_{\text{sub}} g d^3x - I_{h,1} = G_1 + G_2 - I_{h,1}$$

$$I_{e,2} = -q \int_{\text{sub}} g d^3x - I_{e,1} = -G_1 - G_2 - I_{e,1} \quad (4.4a)$$

Taking the gradient of (4.2) and using (2.9) allows the S_1 currents to be expressed in terms of $\text{grad}U$ and $\text{grad}(\Omega+\phi)$, which then allows $I_{h,1}$ to be expressed in terms of $I_{e,1}$ as

$$I_{h,1} = (D_h/D_e) I_{e,1} - 2 q D_h \int_{S_1} \text{grad} (\Omega + \phi) \cdot ds$$

and using (4.3), (2.21), and (2.22) gives

$$I_{h,1} = (V_2 + 2V_T P_2/p_0)/R_0 + (1 + D_h/D_e) G_1 + (D_h/D_e) I_{e,1}$$

and the equation for $I_{h,2}$ becomes

$$I_{h,2} = G_2 - (D_h/D_e) G_1 - (V_2 + 2V_T P_2/p_0)/R_0 - (D_h/D_e) I_{e,1} \quad (4.4b)$$

4.3 Expressing $I_{e,1}$ in Terms of Γ

Using (4.2) to eliminate U in (2.6b) and rearranging terms gives

$$\text{div} \{(P + A) \text{grad} [\]\} = (A_0 - A) \text{div grad } \phi \quad (4.5)$$

where

$$[] \equiv P + (p_0/2 - A) \ln(1 + P/A) - (\Omega + \phi)$$

and A is given by (3.6). Note that A could have been replaced by other constants in the above equations. The motivation for the particular choice A will be clear later. We now define a new unknown function Γ by the boundary value problem

$$\text{div} [(P + A) \text{grad } \Gamma] = 0 \quad \text{in substrate} \quad (4.6a)$$

$$\Gamma = 0 \quad \text{on } S_1 \quad (4.6b)$$

$$\Gamma = 1 \quad \text{on } S_2 \quad (4.6c)$$

The present objective is to express $I_{e,1}$ in terms of Γ , so that an approximation for Γ , which will come later, produces an approximation for $I_{e,1}$. The divergence theorem together with (4.5) and (4.6) gives

$$\begin{aligned} & \int (1 - \Gamma) (P + A) \text{grad} [] \cdot \text{ds} + \int [] (P + A) \text{grad } \Gamma \cdot \text{ds} \\ & + (A - A_0) \int (1 - \Gamma) \text{grad } \phi \cdot \text{ds} = (A_0 - A) \int \text{grad } \Gamma \cdot \text{grad } \phi \, d^3x \end{aligned}$$

where the surface integrals are on both S_1 and S_2 , and the volume integral is over the substrate. Using

$$(P + A) \text{grad} [] = (P + p_0/2) \text{grad } P - (P + A) \text{grad} (\Omega + \phi)$$

together with (4.6b) and (4.6c) gives

$$\begin{aligned}
(p_0/2) \int_{S_1} \text{grad } P \cdot \mathbf{ds} &= A \int_{S_1} \text{grad } (\Omega + \phi) \cdot \mathbf{ds} - \int [] (P+A) \text{grad } \Gamma \cdot \mathbf{ds} \\
&+ (A_0 - A) \left[\int_{S_1} \text{grad } \phi \cdot \mathbf{ds} + \int_{\text{sub}} \text{grad } \Gamma \cdot \text{grad } \phi \, d^3x \right]
\end{aligned}$$

which expresses the left side in terms of known quantities and the unknown Γ . The motivation for selecting A instead of some other constant is that $[] = 0$ on S_1 and S_2 . The equation reduces to

$$\begin{aligned}
(p_0/2) \int_{S_1} \text{grad } P \cdot \mathbf{ds} &= A \int_{S_1} \text{grad } (\Omega + \phi) \cdot \mathbf{ds} \\
&+ (A_0 - A) \left[\int_{S_1} \text{grad } \phi \cdot \mathbf{ds} + \int_{\text{sub}} \text{grad } \Gamma \cdot \text{grad } \phi \, d^3x \right]
\end{aligned}$$

so the unknown Γ appears only in a weight factor in a weighted average. This observation will be used in the next section, which produces an approximation for $I_{e,1}$. The above equation can be expressed in terms of $I_{e,1}$ using (2.9) and (4.2) with the result

$$\begin{aligned}
[p_0/(2 \, q \, D_e)] I_{e,1} &= A \int_{S_1} \text{grad } (\Omega + \phi) \cdot \mathbf{ds} \\
&+ (A_0 - A) \left[\int_{S_1} \text{grad } \phi \cdot \mathbf{ds} + \int_{\text{sub}} \text{grad } \Gamma \cdot \text{grad } \phi \, d^3x \right] \cdot (4.7)
\end{aligned}$$

4.4 An Approximation for P and the Currents

The role that Γ plays in (4.7) is most visible when the equation is written in one dimension as...

$$[p_0/(2 q D_e)] I_{e,1} \text{ (per unit area)} = - A d(\Omega+\phi)/dx_0 \\ + (A_0 - A) \left[- d\phi/dx_0 + \int_0^L (d\Gamma/dx) (d\phi/dx) dx \right] \quad (1 \text{ dim.}) \quad (4.8)$$

where S_1 is at $x=0$, S_2 is at $x=L$, and d/dx_0 is abbreviated notation for the derivative evaluated at $x=0$. The normalization condition (4.6b) and (4.6c) can be written as

$$\int_0^L (d\Gamma/dx) dx = 1$$

so $d\Gamma/dx$ in (4.8) is the weight factor in a weighted average of $d\phi/dx$. Integrating (4.6) gives an alternate expression for the weight factor

$$d\Gamma/dx = \left[\int_0^L 1/(P + A) dx \right]^{-1} [1/(P + A)] \quad (1 \text{ dim.}) \quad (4.9)$$

If V_2 is positive and large, A is very small and the weight factor is concentrated near $x=0$, where $P=0$. The weighted average reduces to the endpoint value at $x=0$ and $I_{e,1}$ is small. This is the expected result when V_2 is large.

Weighted averages are usually insensitive to small errors in the weight factor, and this suggests that $I_{e,1}$ can be approximated by replacing the unknown Γ in (4.7) with an approximation. An approximation for Γ is obtained by replacing the unknown P in

(4.6) with an approximation. For LILC, we can assume that $P \ll A$, so it does not matter how we approximate P in (4.6), as long as the approximation is consistent with $P \ll A$. We therefore look for an approximation applicable to high-injection-level conditions (HILC).

The present objective is to find an approximation for P applicable when $P \gg p_0/2$ throughout most of the substrate. A tentative approximation is P^∇ which is defined by

$$P^\nabla + (p_0/2 - A) \ln(1 + P^\nabla/A) = \Omega + \phi \quad (4.10)$$

and satisfies the required boundary conditions. To establish the credibility of the approximation P^∇ , note that (4.5) can be manipulated into

$$\text{div} [(P + p_0/2) \text{grad } P] = \text{div} [(P + A_0) \text{grad } (\Omega + \phi)] \quad (4.11a)$$

while (4.10) can be used to show that

$$\text{div} [(P^\nabla + p_0/2) \text{grad } P^\nabla] = \text{div} [(P^\nabla + A) \text{grad } (\Omega + \phi)] \quad (4.11b)$$

The two equations differ only in that one contains A_0 while the other contains A . The constant A_0 is on the order of $p_0/2$, while A will be of the same order or smaller. For HILC, we will have $P \gg A, A_0$ throughout most of the substrate; it is reasonable to assume that the A 's have little influence, i.e., $P \approx P^\nabla$. Note that if the approximation works at all, it is not limited to locations where P is large. The right sides of (4.11) can be thought of as driving terms, analogous to charge density, which have accumulating effects in the sense that the solution anywhere is influenced by the charge density everywhere. If the charge densities are nearly equal throughout most of the substrate, where they are greatest, the solutions will be nearly equal everywhere. If $P \gg p_0/2$ throughout most of the substrate, so that $P \approx P^\nabla$ throughout most of the substrate, we will also have $P \approx P^\nabla$ near S_1 where P is small.

Quantitative calculations of P^∇ can be done by using the special function F (discussed in Appendix C) to write (4.10) as

$$P^\nabla = (p_0/2) F(2A/p_0, 2(\Omega+\phi)/p_0) \quad (4.12)$$

but approximations are useful for visualization. Note that (4.12) and (3.11) are the same except that Ω is replaced by $\Omega+\phi$. The generalized ambipolar approximation is obtained by making the same replacement. Neglecting $(p_0/2V_T)V_2$ compared to P_2 for the negative V_2 case, the approximation is

$$P^\nabla \approx \Omega + \phi \quad \text{if } V_2 \leq 0 \quad \text{and} \quad P_2 \gg p_0/2. \quad (4.13)$$

If $V_2 > 0$, there is an AR and HRR separated by an ARB, which is the constant $\Omega+\phi$ surface characterized by

$$\Omega + \phi = (p_0/2V_T) V_2 \quad \text{defines ARB} . \quad (4.14)$$

The approximation in the AR is

$$P^\nabla \approx \Omega + \phi - (p_0/2V_T) V_2 \quad \text{in AR if } V_2 > 0 \quad \text{and} \quad P_2 \gg p_0/2 . \quad (4.15)$$

Quantitative estimates of P in the HRR (and anywhere else) should use (4.12); but, for visualization purposes, it is enough to know that P^∇ is much smaller in the HRR than in the AR.

Returning to $I_{e,1}$, the approximation is obtained by replacing Γ in (4.7) with Γ^∇ defined by

$$\text{div} [(P^\nabla + A) \text{grad } \Gamma^\nabla] = 0 \quad \text{in substrate} \quad (4.16a)$$

$$\Gamma^\nabla = 0 \quad \text{on } S_1 \quad (4.16b)$$

$$\Gamma^\nabla = 1 \quad \text{on } S_2 . \quad (4.16c)$$

With surface integrals expressed in terms of R_0 and G_1 , and A_0 related to D^* , the approximation can be written as

$$I_{e,1} \approx - (D_e/D_h) (A/p_0) (V_2 + 2V_T P_2/p_0)/R_0 - G_1 \\ + 2D_e q [(A_0 - A)/p_0] \int_{\text{sub}} \text{grad } \Gamma^\nabla \cdot \text{grad } \phi \, d^3x \quad (4.17a)$$

Currents at S_2 are estimated by substituting the above result into (4.4) to get

$$I_{h,2} \approx - (1 - A/p_0) (V_2 + 2V_T P_2/p_0)/R_0 + G_2 \\ - 2D_h q [(A_0 - A)/p_0] \int_{\text{sub}} \text{grad } \Gamma^\nabla \cdot \text{grad } \phi \, d^3x \quad (4.17b)$$

$$I_{e,2} \approx (D_e/D_h) (A/p_0) (V_2 + 2V_T P_2/p_0)/R_0 - G_2 \\ - 2D_e q [(A_0 - A)/p_0] \int_{\text{sub}} \text{grad } \Gamma^\nabla \cdot \text{grad } \phi \, d^3x \quad (4.17c)$$

The equations in (4.17) are approximations, but the particular combination of equations given by

$$(D_h/D_e)I_{e,2} - I_{h,2} = (V_2 + 2V_T P_2/p_0)/R_0 - (1 + D_h/D_e) G_2 \quad (4.18)$$

is exact.

4.5 A Mathematical Theorem

The integral in (4.17) has an interpretation (as a weighted average of $\text{grad}\phi$), but is difficult to numerically evaluate in three dimensions. The objective of this and the remaining sections is to make (4.17) computationally manageable. The first step towards this objective is to derive a theorem relating volume integrals to surface integrals. The identity derived here is a little more versatile, for our applications, than the usual divergence theorem.

Let $S(v)$ denote the constant Ω_u surface characterized by $\Omega_u=v$. Note that v can be used as one coordinate in a curvilinear coordinate system. The value of v determines which constant Ω_u surface a given space point lies on. Let τ_1 and τ_2 be two surface coordinates selected so that (τ_1, τ_2, v) form an orthogonal system. If \mathbf{J} is a sectionally continuous, but otherwise arbitrary vector field, we have

$$\int_{\text{sub}} \mathbf{J} \cdot \text{grad } \Omega_u \, d^3x = \int \int \int \mathbf{J} \cdot \text{grad } \Omega_u \, h_1 \, h_2 \, h_3 \, d\tau_1 \, d\tau_2 \, dv$$

where h_1 , h_2 , and h_3 are the scale factors for the coordinates τ_1 , τ_2 , and v , respectively. But h_3 is given by [4]

$$h_3 = |\text{grad } \Omega_u|^{-1}$$

so the equation becomes

$$\int_{\text{sub}} \mathbf{J} \cdot \text{grad } \Omega_u \, d^3x = \int_0^1 \left[\int \int \mathbf{J} \cdot \mathbf{n} \, h_1 \, h_2 \, d\tau_1 \, d\tau_2 \right] dv$$

where \mathbf{n} is the unit vector in the direction of $\text{grad}\Omega_u$. The double integral inside of the brackets is a surface integral on the $\Omega_u=v$ surface, so the equation now becomes

$$\int_{\text{sub}} \mathbf{J} \cdot \text{grad } \Omega_u d^3x = \int_0^1 \int_{S(v)} \mathbf{J} \cdot d\mathbf{s} dv \quad (\text{arbitrary } \mathbf{J}). \quad (4.19a)$$

The normal unit vector in the surface integral is in the direction of increasing Ω_u , so it is directed outward from the region between S_1 and $S(v)$. We therefore have

$$\int_{S(0)} \mathbf{J} \cdot d\mathbf{s} = - \int_{S_1} \mathbf{J} \cdot d\mathbf{s}, \quad \int_{S(1)} \mathbf{J} \cdot d\mathbf{s} = \int_{S_2} \mathbf{J} \cdot d\mathbf{s}.$$

A trivial generalization of the above steps gives

$$\int_{R(v)} \mathbf{J} \cdot \text{grad } \Omega_u d^3x = \int_0^v \int_{S(v')} \mathbf{J} \cdot d\mathbf{s} dv' \quad (4.19b)$$

where $R(v)$ is the region between S_1 and $S(v)$.

4.6 A Special Family of Generation Rate Functions

The second step towards the goal of making (4.17) computationally manageable is to confine our attention to a special family of generation rate functions. It will be assumed that g can be expressed in the form

$$g = \alpha(\Omega_u) \text{grad } \Omega_u \cdot \text{grad } \Omega_u \quad (4.20)$$

for some function α . It is always possible to express g in the form (4.20) in one dimension because the product of the gradients is a constant and the argument of α is a linear function of the spatial coordinate. If the substrate has length L and we are given a $g(x)$ with the origin selected so that S_1 is at $x=0$ and S_2 is at $x=L$, then $\Omega_u=x/L$ and $\alpha(v)=L^2g(vL)$. But (4.20) imposes a

restriction in three dimensions. If we are selecting a g to represent a hypothetical case of our own choice, we can always select it to have the form (4.20). A more probable situation is one in which a g has been given and there is no α satisfying (4.20). We then look for an α that gives some kind of best fit, or at least a good fit (if possible). It is left to the user to find a fitting function α , but some guidance is given below.

Selection of a fitting function α may be a little easier if α is related to familiar physical quantities. Such quantities are G_1 , G_2 , and the volume integral of g . We start with

$$\phi = (\Omega_u/D^*) \int_0^1 \int_0^v \alpha(v_1) dv_1 dv - (1/D^*) \int_0^{\Omega_u} \int_0^v \alpha(v_1) dv_1 dv \quad (4.21)$$

which can be verified by substituting (4.20) and (4.21) into (2.20). The gradient is given by

$$D^* \text{grad } \phi = \left[\int_0^1 \int_0^v \alpha(v_1) dv_1 dv - \int_0^{\Omega_u} \alpha(v) dv \right] \text{grad } \Omega_u \quad (4.22)$$

so

$$\begin{aligned} G_2 &= -q D^* \int_{S_2} \text{grad } \phi \cdot ds \\ &= V_T (D_h P_o R_o)^{-1} \left[\int_0^1 \alpha(v) dv - \int_0^1 \int_0^v \alpha(v_1) dv_1 dv \right] \end{aligned} \quad (4.23a)$$

where we have used (2.21). Similarly,

$$G_1 = V_T (D_h P_o R_o)^{-1} \int_0^1 \int_0^v \alpha(v_1) dv_1 dv \quad (4.23b)$$

and

$$q \int_{\text{sub}} g d^3x = G_1 + G_2 = V_T (D_h P_O R_O)^{-1} \int_0^1 \alpha(v) dv . \quad (4.23c)$$

The three equations (4.23) relate α to familiar physical quantities and may provide some guidance for those looking for a fitting function α (one good method is derived in Section 6.4). But the analysis given here goes in the other direction. It is assumed that α has been provided and the objective is to calculate other quantities from it. When going in this direction, it is convenient to express quantities in terms of β instead of α , where β is defined by

$$D^* (P_O/2) \beta(v) \equiv v \int_0^1 \int_0^{v_2} \alpha(v_1) dv_1 dv_2 - \int_0^v \int_0^{v_2} \alpha(v_1) dv_1 dv_2 \quad (4.24a)$$

so that

$$D^* (P_O/2) \beta'(v) = \int_0^1 \int_0^{v_2} \alpha(v_1) dv_1 dv_2 - \int_0^v \alpha(v_1) dv_1 . \quad (4.24b)$$

The only thing that we need α for is to construct β and β' . The latter functions will be used from now on. Combining (4.24) with the previous equations gives

$$\phi = (P_O/2) \beta(\Omega_U) \quad (4.25)$$

$$G_1 = [D_e / (D_e + D_h)] (V_T / R_O) \beta'(0) \quad (4.26a)$$

$$G_2 = - [D_e / (D_e + D_h)] (V_T / R_O) \beta'(1) . \quad (4.26b)$$

Another important quantity is the sum $\Omega + \phi$ which is expressed as

$$\Omega + \phi = (p_0/2) \beta_m(\Omega_u) \quad (4.27)$$

where β_m is a modified β defined by

$$\beta_m(v) \equiv \beta(v) + (v_2/v_T + 2P_2/p_0) v \quad (4.28)$$

and is trivially related to β . A separate symbol is used only for notational brevity. We can write (4.12) in terms of β_m as

$$P^\nabla = (p_0/2) F(2A/p_0, \beta_m(\Omega_u)) .$$

For notational brevity, we will leave out the first argument and write the equation as

$$P^\nabla = (p_0/2) F(\beta_m(\Omega_u)) \quad (\text{abbreviated notation}) . \quad (4.29)$$

The integral in (4.17) can be evaluated by using (4.19) together with $\text{grad}\phi = (p_0/2)\beta'(\Omega_u)\text{grad}\Omega_u$ to get

$$\int_{\text{sub}} \text{grad } \Gamma^\nabla \cdot \text{grad } \phi \, d^3x = (p_0/2) \int_0^1 \beta'(v) \int_{S(v)} \text{grad } \Gamma^\nabla \cdot \, ds \, dv . \quad (4.30)$$

But

$$\Gamma^\nabla = \frac{\int_0^{\Omega_u} [(p_0/2) F(\beta_m(v)) + A]^{-1} \, dv}{\int_0^1 [(p_0/2) F(\beta_m(v)) + A]^{-1} \, dv} \quad (4.31)$$

which can be verified by substituting (4.29) and (4.31) into (4.16). Taking the gradient of (4.31) and substituting it into (4.30) while using

$$\int_{S(v)} \text{grad } \Omega_u \cdot ds = \int_{S_2} \text{grad } \Omega_u \cdot ds = V_T (q D_h p_o R_o)^{-1}$$

gives

$$\int_{\text{sub}} \text{grad } \Gamma^v \cdot \text{grad } \phi d^3x = V_T (2q D_h R_o)^{-1} (\text{INT2}/\text{INT1}) \quad (4.32)$$

with the two integrals INT1 and INT2 defined by

$$\text{INT1} \equiv \int_0^1 [(p_o/2) F(\beta_m(v)) + A]^{-1} dv \quad (4.33a)$$

$$\text{INT2} \equiv \int_0^1 [(p_o/2) F(\beta_m(v)) + A]^{-1} \beta'(v) dv \quad (4.33b)$$

The ratio INT2/INT1 is a weighted average of β' , similar to the weighted average of $d\phi/dx$ in the one-dimensional equations (4.8) and (4.9).

The currents are estimated by substituting (4.26) and (4.32) into (4.17) to get

$$\begin{aligned} I_{h,2} \approx & - (1 - A/p_o) (V_2 + 2V_T P_2/p_o)/R_o \\ & - (D_e/D_h) (A_o/p_o) (V_T/R_o) \beta'(1) \\ & - [(A_o - A)/p_o] (V_T/R_o) (\text{INT2}/\text{INT1}) \end{aligned} \quad (4.34a)$$

$$\begin{aligned}
I_{e,2} \approx & (D_e/D_h) (A/p_o) (V_2 + 2V_T P_2/p_o)/R_o \\
& + (D_e/D_h) (A_o/p_o) (V_T/R_o) \beta'(1) \\
& - (D_e/D_h) [(A_o - A)/p_o] (V_T/R_o) (INT2/INT1) \quad (4.34b)
\end{aligned}$$

$$(D_h/D_e)I_{e,2} - I_{h,2} = (V_2 + 2V_T P_2/p_o)/R_o + (V_T/R_o)\beta'(1) \quad (4.35)$$

The two equations (4.34) are approximations while (4.35) is exact. Any two of the above three equations can be used to solve for the currents.

4.7 A Numerical Integration

With a function β given, all quantities on the right sides of (4.34) can be calculated, but the integrals INT1 and INT2 given by (4.33) require numerical methods. The numerical integration is regarded as part of the theory, rather than an exercise left for the reader, so some discussion is given here.

The reader might notice that some of the integration can be done analytically. The derivatives β' and β_m' differ by a constant, so both integrals can be evaluated if we can evaluate INT1 and the integral

$$\int [(p_o/2) F(\beta_m(v)) + A]^{-1} \beta_m'(v) dv = \int [(p_o/2)F + A]^{-1} (d\beta_m/dF) dF .$$

$F(\beta_m)$ is related to β_m by

$$F + (1 - 2A/p_o) \ln(1 + p_o F/2A) = \beta_m$$

which allows $d\beta_m/dF$ to be expressed in terms of F alone. The above integral can be expressed in closed form, so only INT1

requires numerical approximations. This method is intentionally not used, because it is equivalent to approximating Γ^∇ in (4.31) by retaining the numerator on the right while approximating the denominator with a numerical estimate. Any error in the estimate upsets the normalization condition $\Gamma^\nabla(1)=1$. The estimates of the currents are insensitive to errors in Γ or in Γ^∇ when properly normalized, but estimates are sensitive to errors that disturb the normalization. If this method is used, accurate current estimates require an accurate numerical estimate of INT1. This is not easy, because the integrand can be extremely skewed, requiring a carefully selected variable step size for accurate numerical approximation. It is desirable to eliminate the need for such numerical sophistication by using a different method to evaluate the integrals.

One simple method is to numerically approximate both integrals, using the same step sizes for both. To see why this works, note that the ratio INT2/INT1 is a weighted average of β' . Even if the weight factor is extremely skewed, the step size need be no larger than dictated by β' (i.e., the step size only needs to be small enough for β' to be nearly constant in each subinterval) if the numerical approximation of the weight factor is correspondingly skewed and normalized. By using the same step sizes for both integrals, we insure that the numerical approximation of the weight function is normalized, even if the step sizes are not small enough for an accurate estimate of INT1. We can therefore use a uniform step size to evaluate the integrals.

One potential source of numerical error, which gets worse with smaller step sizes, can and should be avoided. This error source is the subtraction of nearly equal numbers that will occur when using $\beta'dv=d\beta$. It is better to leave $\beta'dv$ as it is. This means that the user is required to supply β' in addition to β , but this is not a lot of extra work. If the user can calculate β from (4.24a), then the user can also calculate β' from (4.24b).

A suggested numerical integration is the following. Select a moderately large value for M (the numerical examples in Chapter 6 used $M=100$) and then calculate the quantities listed below (arrays are obviously unnecessary if the quantities are calculated when needed):

$$X_1 = 2A/p_0$$

$$B_i' = \beta'(i/M) \quad i=0, \dots, M$$

$$B_i = \beta(i/M) + (V_2/V_T + 2P_2/p_0) (i/M) \quad i=0, \dots, M$$

$$C_i = [(p_0/2) F(X_1, B_i) + A]^{-1} \quad i=0, \dots, M$$

$$\text{INT1} \approx (C_0 + C_M)/(2M) + (1/M) \sum_{i=1}^{M-1} C_i$$

$$\text{INT2} \approx (C_0 B_0' + C_M B_M')/(2M) + (1/M) \sum_{i=1}^{M-1} C_i B_i'$$

5. THE COMPLETE SOLUTION

5.1 Introduction

This chapter does little more than list the equations in Chapter 4 together with those in Appendix A, to produce a complete equation set that is able to solve for all currents and boundary values. The only effort required here is associated with nuisance details such as including an electrode-semiconductor contact potential, and selecting a notation common to both equation sets. Following the list of equations is a suggested algorithm for constructing device I-V curves. This algorithm is interpreted as the "complete solution." A simple necessary condition for saturation is derived in the last section.

5.2 Notation

The notation used for the substrate analysis is familiar by now and the notation used for the DR analysis is listed in Appendix A. Redundant notations are related below so that the redundancy can be eliminated. The scalar current densities in the DR equations are evaluated at the DRB on the lightly doped side, which is S_2 . These currents are positive when directed from the n-side towards the p-side, so

$$I_{h,2} = -j_h A_D, \quad I_{e,2} = -j_e A_D \quad \text{for the p-type substrate}$$

$$I_{h,2} = j_h A_D, \quad I_{e,2} = j_e A_D \quad \text{for the n-type substrate}$$

where A_D is the DRB surface area. The total current I is also taken to be positive when directed from the n-side towards the p-side, so

$$I_{T,2} = -I \equiv -j_T A_D \quad \text{for the p-type substrate} \quad (5.1a)$$

$$I_{T,2} = I \equiv j_T A_D \quad \text{for the n-type substrate} \quad (5.1b)$$

$$j_T \equiv j_h + j_e \quad (5.1c)$$

The equilibrium majority carrier density is equated to the doping density, so

$$p_0 = N_A \quad \text{for the p-type substrate}$$

$$n_0 = N_D \quad \text{for the n-type substrate.}$$

The equilibrium minority carrier density was left out of the substrate equations, but retained in some of the DR equations. We therefore use

$$n_p = P_2 + n_0 \quad \text{where } n_0 = n_i^2/N_A \quad \text{for the p-type substrate}$$

$$p_n = P_2 + p_0 \quad \text{where } p_0 = n_i^2/N_D \quad \text{for the n-type substrate}$$

where n_i is the intrinsic electron density.

Contact potentials between electrodes and semiconductor are simulated by fictitious power supplies of voltage V_C as shown in Figure 5.1. The p- and n-type substrates are both shown. In each case, the polarity of the fictitious power supply is chosen so that V_C is positive. V_C is given by the well-known equation

$$V_C = V_T \ln(N_A N_D/n_i^2) . \quad (5.2)$$

Lumped resistors R_C (Figure 5.1) simulate ohmic contact resistances, and may also include any other desired circuit resistances associated with electrical connections outside of the diode interior. The voltage V is applied to the upper contact (Figure 5.1) above any resistor elements that are included in R_C . A current arrow indicates the direction of the current when I is positive, consistent with the sign convention stated above. The potentials are related by

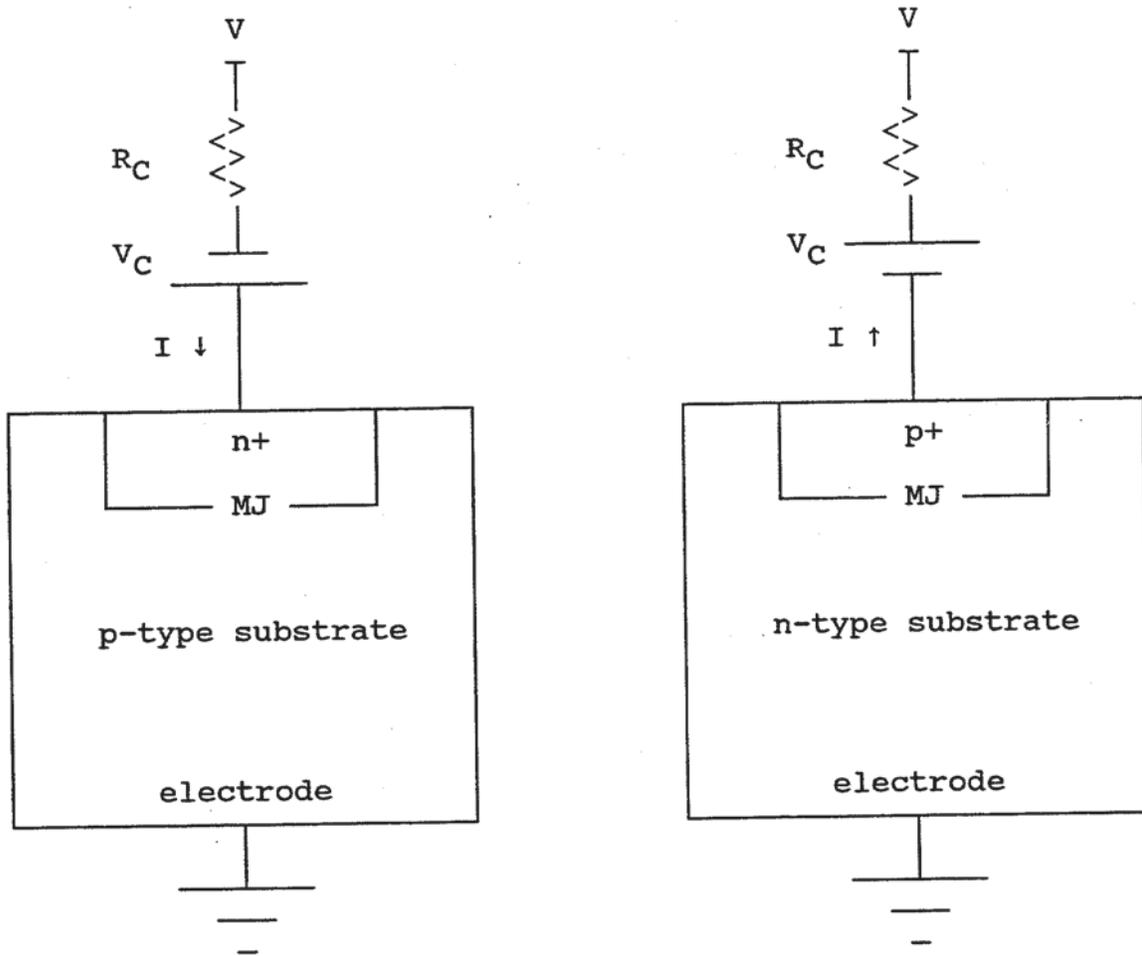


Figure 5.1: Qualitative sketch of both diode types showing R_C and V_C . The currents are positive when in the indicated directions.

$$V = V_2 + V_{DR} - V_C + I R_C \quad \text{for p-type substrate} \quad (5.3a)$$

$$V = V_2 - V_{DR} + V_C - I R_C \quad \text{for n-type substrate.} \quad (5.3b)$$

5.3 Equation Summary for the n⁺/p Diode

All equations, excluding those listed in Section 5.2 and geometric information that must be supplied by the reader, are listed here for the p-type substrate diode.

Starting with the doping densities N_A (p-side) and N_D (n-side), the low field mobilities $\mu_{o,h}$ and $\mu_{o,e}$, the saturation velocity v , the thermal voltage V_T , the elementary charge q , and the dielectric constant ϵ , other constants are calculated from

$$D_h = V_T \mu_{o,h} , \quad D_e = V_T \mu_{o,e} , \quad D^* = 2D_h D_e / (D_h + D_e) \quad (5.4a)$$

$$\sigma_o = (q/V_T) D_h N_A , \quad A_o = D_h N_A / (D_h + D_e) \quad (5.4b)$$

$$a_e = 1/(q V_T \mu_{o,e}) , \quad V_T b = 1/(q v) . \quad (5.4c)$$

Boundary values P_2 and V_2 must be solved. The parameters A and E are defined in terms of P_2 and V_2 by

$$E = [\ln(1 + 2P_2/N_A)]^{-1} \quad \text{if } V_2 = 0 \quad \text{and } P_2 > 0 . \quad (5.5a)$$

If $V_2 \neq 0$ and $V_2 \neq -2V_T P_2/N_A$, use

$$Z_1 \equiv (V_T/V_2) (1 + 2P_2/N_A) , \quad Z_2 \equiv V_T/V_2 \quad (5.5b)$$

$$E = H(Z_1, Z_2) \quad (5.5c)$$

where the special function H is defined in Appendix B. For any

case such that $V_2 \neq -2V_T P_2 / N_A$, use either

$$A = (N_A/2) [1 - (V_2/V_T) E] \quad (5.5d)$$

or

$$A = P_2 e^{-1/E} (1 - e^{-1/E})^{-1} . \quad (5.5e)$$

The two equations for A give the same result in theory, but the second should be used if $(V_2/V_T)E$ is so nearly equal to 1 that the first requires more numerical precision than is available. Otherwise, the first can be used. The functions Ω_u , ϕ , and Ω are defined by

$$\text{div grad } \Omega_u = 0 \text{ in sub. , } \Omega_u = 0 \text{ on electrode, } \Omega_u = 1 \text{ on DRB}$$

$$\text{div grad } \phi = -g/D^* \text{ in sub. , } \phi = 0 \text{ on electrode, } \phi = 0 \text{ on DRB}$$

$$\Omega = [P_2 + (N_A/2V_T) V_2] \Omega_u$$

with reflective boundary conditions on the insulated boundaries tacitly assumed. The electrical resistance between electrode and DRB produced by the uniform conductivity σ_0 is R_0 . The ambipolar diffusion currents G_1 and G_2 are given by

$$G_i = -q D^* \int_{S_i} \text{grad } \phi \cdot ds \quad (i = 1, 2)$$

with the unit normal vector chosen so that G_i is positive. R_0 and the G 's may depend on the DR width W .

An approximation for P applicable when $P \gg N_A/2$ throughout most of the substrate is P^∇ given by

$$P^\nabla = (N_A/2) F(2A/N_A, 2(\Omega+\phi)/N_A)$$

where the special function F is discussed in Appendix C. The simpler generalized ambipolar approximation is useful for visualization when $P_2 \gg N_A/2$. If $V_2 \leq 0$, the approximation is

$$P^\nabla \approx \Omega + \phi \quad \text{if } V_2 \leq 0 \quad \text{and} \quad P_2 \gg N_A/2.$$

If $V_2 > 0$, there is an AR and HRR separated by an ARB, which is the constant $\Omega + \phi$ surface characterized by

$$\Omega + \phi = (N_A/2V_T) V_2 \quad \text{defines ARB} . \quad (5.6)$$

P^∇ is small in the HRR, but the approximation in the AR is

$$P^\nabla \approx \Omega + \phi - (N_A/2V_T) V_2 \quad \text{in AR if } V_2 > 0 \quad \text{and} \quad P_2 \gg N_A/2 .$$

Approximations for the currents are obtained by first defining Γ^∇ by

$$\text{div} [(P^\nabla + A) \text{grad } \Gamma^\nabla] = 0 \quad \text{in substrate}$$

$$\Gamma^\nabla = 0 \quad \text{on electrode,} \quad \Gamma^\nabla = 1 \quad \text{on DRB} .$$

The currents are approximated by

$$j_h A_D \approx (1 - A/N_A) \cdot (V_2 + 2V_T P_2/N_A)/R_0 - G_2 \\ + 2D_h q [(A_0 - A)/N_A] \int_{\text{sub}} \text{grad } \Gamma^\nabla \cdot \text{grad } \phi \, d^3x$$

$$(D_h/D_e)j_e = j_h + (1 + D_h/D_e)G_2/A_D - (V_2 + 2V_T P_2/N_A)/(A_D R_0) . \quad (5.7)$$

Calculations are manageable in three dimensions if g can be expressed as

$$g = \alpha(\Omega_u) \text{grad } \Omega_u \cdot \text{grad } \Omega_u \quad (5.8)$$

for some function α , which is used to construct the user-supplied function β and derivative β' given by

$$D^*(N_A/2) \beta(v) = v \int_0^1 \int_0^{v_2} \alpha(v_1) dv_1 dv_2 - \int_0^v \int_0^{v_2} \alpha(v_1) dv_1 dv_2 \quad (5.9a)$$

$$D^*(N_A/2) \beta'(v) = \int_0^1 \int_0^{v_2} \alpha(v_1) dv_1 dv_2 - \int_0^v \alpha(v_1) dv_1 \quad (5.9b)$$

The modified β is given by

$$\beta_m(v) = \beta(v) + (V_2/V_T + 2P_2/N_A) v \quad (5.9c)$$

so that

$$\Omega + \phi = (N_A/2) \beta_m(\Omega_u) \quad (5.10)$$

$$P^\nabla = (N_A/2) F(2A/N_A, \beta_m(\Omega_u)) \quad (5.11)$$

The currents are now approximated by

$$\begin{aligned} j_h \approx & [(N_A - A)/N_A] (V_2 + 2V_T P_2/N_A) / (A_D R_O) \\ & + V_T [(N_A - A_O)/N_A] \beta'(1) / (A_D R_O) \\ & + V_T [(A_O - A)/N_A] (\text{INT2}/\text{INT1}) / (A_D R_O) \end{aligned} \quad (5.12a)$$

$$(D_h/D_e) j_e = j_h - [V_2 + V_T \beta'(1) + 2V_T P_2/N_A]/(A_D R_0) \quad (5.12b)$$

where the two integrals INT1 and INT2 are evaluated by selecting a moderately large M (e.g., 100) and using

$$X_1 = 2A/N_A \quad (5.13a)$$

$$B_i' = \beta'(i/M) \quad i=0, \dots, M \quad (5.13b)$$

$$B_i = \beta(i/M) + (V_2/V_T + 2P_2/N_A) (i/M) \quad i=0, \dots, M \quad (5.13c)$$

$$C_i = [(N_A/2) F(X_1, B_i) + A]^{-1} \quad i=0, \dots, M \quad (5.13d)$$

$$INT1 \approx (C_0 + C_M)/(2M) + (1/M) \sum_{i=1}^{M-1} C_i \quad (5.13e)$$

$$INT2 \approx (C_0 B_0' + C_M B_M')/(2M) + (1/M) \sum_{i=1}^{M-1} C_i B_i' \quad (5.13f)$$

Note that (5.12b) can be rewritten as

$$V_2 = A_D R_0 (j_h - D_h j_e/D_e) - 2V_T P_2/N_A - V_T \beta'(1) \quad (5.14)$$

One of the DR equations is

$$j_h = q W g_D \quad (5.15)$$

with W the DR width and g_D the value of g at the DR location. Another DR equation is

$$\begin{aligned} \exp(-V_{DR}/V_T) &= N_D^{-1} [P_2 + n_o - V_{Tb} j_T] \\ &\quad - N_D^{-1} (V_T \epsilon/q) (a_e j_T)^2 [N_A \\ &\quad + V_{Tb} j_T]^{-1} [N_A - V_{Tb} j_T + 2P_2 + 2n_o]^{-1} \quad \text{if } j_T > 0 \end{aligned}$$

$$\exp(-V_{DR}/V_T) = N_D^{-1} (P_2 + n_o) \quad \text{if } j_T \leq 0$$

which can be solved for P_2 in terms of j_T and the DR voltage drop V_{DR} using

$$P_2 = N_D \exp(-V_{DR}/V_T) - n_o \quad \text{if } j_T \leq 0 \quad (5.16a)$$

$$T_1 \equiv N_D \exp(-V_{DR}/V_T) + V_{Tb} j_T \quad (5.16b)$$

$$T_2 \equiv (V_T/2) (\epsilon/q) (a_e j_T)^2 / (N_A + V_{Tb} j_T) \quad (5.16c)$$

$$T_3 \equiv N_D \exp(-V_{DR}/V_T) + (1/2) V_{Tb} j_T + N_A/2 \quad (5.16d)$$

$$T_4 \equiv T_3 + [T_3^2 + 4T_2]^{1/2} \quad (5.16e)$$

$$P_2 = T_1 - n_o + 2T_2/T_4 \quad \text{if } j_T > 0 \quad (5.16f)$$

The DR equation used to solve for W is

$$\begin{aligned} W &= (2\epsilon/q)^{1/2} V_{DR}^{1/2} [(N_A + V_{Tb} j_T)^{v6/2} \\ &\quad + (2\epsilon/q)^{1/\sqrt{6}} (V_T a_e j_T)^{v6/3} V_{DR}^{-1/\sqrt{6}}]^{-1/\sqrt{6}} \quad \text{if } j_T > 0 \quad (5.17a) \end{aligned}$$

$$W = [(2\epsilon/q) V_{DR}/N_A]^{1/2} \quad \text{if } j_T \leq 0 \quad (5.17b)$$

5.4 Algorithm for Constructing the n^+/p Diode I-V Curve

A suggested algorithm for constructing I-V curves for the p-type substrate diode is listed below. The voltage polarity and direction of current when positive are shown in Figure 5.1. An example of an I-V curve is seen by looking ahead to Figure 6.1 in Chapter 6. The diode delivers power (solar cell operation) when V is negative (a forward-biasing polarity) with I positive (a reverse current produced by photogeneration). The "model" curve for the particular example shown in the figure saturates for V greater than about -0.4 volts. Numerical problems will result if we try to extend the curve too far into saturation, because A calculated from (5.5) becomes so close to zero that finite numerical precision fails to distinguish it from zero. But there is no need to extend the plot beyond the point where such a problem first occurs, because such a point is far into saturation. In the opposite extreme of small (negative) V, the curve is very steep. Attempting to extend the curve too far in this direction also produces numerical problems because some calculated quantities become extremely sensitive to tiny errors (smaller than machine precision) in other quantities. But there is no need to extend the plot beyond the point where such problems begin to occur, because the current is large enough (in absolute value) to destroy the device. The objective is to plot points in the "range of interest," which is the range that avoids numerical problems and should also be the range that is physically interesting. A suggested algorithm is the following:

(1) Assign values to q , ϵ/q , V_T , N_A , N_D , n_0 ($=n_i^2/N_A$), V_C (using (5.2)), R_C , A_D , g_D , and the constants on the left sides of (5.4).

(2) Select a positive value for V_{DR} . Each selected value will produce one point on the I-V curve. Trial and error is the simplest way to find a V_{DR} value that produces a point in the range of interest. After several I-V points have been plotted, they can guide later selections of V_{DR} values.

(3) Guess at a value for j_T .

(4) Use (5.16) to solve for P_2 . Change the value to zero if the presence of n_0 in (5.16) produces a negative value.

(5) Use (5.17) to solve for W and (5.15) to solve for j_h .

Then calculate j_e from $j_T - j_h$.

(6) With a value assigned to the DR width W , the substrate geometry is also specified. Assign a value to R_o . Find a fitting function α that (approximately) satisfies (5.8), and use (5.9) to construct the functions β and β' .

(7) Use (5.14) to solve for V_2 .

(8) Use (5.5) to solve for E . The function subprogram in Appendix B can be appended to any FORTRAN driver code for numerical evaluation of the function H . Note that the computer version of H contains a redundant argument Z_3 for improved numerical accuracy. Before calculating E , first calculate Z_3 from $Z_3 = 2(V_T/V_2)(P_2/N_A)$. Then calculate E from $E = H(Z_1, Z_2, Z_3)$.

(9) Use (5.5) to solve for A . If A is found to be negative, the j_T guess was probably too large. Try a less positive or a more negative j_T . If A is positive but so close to zero that the available numerical precision cannot distinguish it from zero when (5.5e) is used, it is probable that either the j_T guess was too small, or the V_{DR} selection places the I-V point too far into saturation. First try a larger j_T . If convergence (step 12 below) cannot be obtained with j_T large enough to avoid this problem, use a smaller V_{DR} .

(10) Use (5.13) to calculate the integrals INT1 and INT2. The function subprogram in Appendix C can be appended to any FORTRAN driver code for numerical evaluation of the function F .

(11) Use (5.12a) to calculate a new value for j_h , denoted $j_{h,new}$. Then calculate $\delta j_h = j_{h,new} - j_h$. Calculate I from $j_T A_D$ and then use (5.3) to calculate V .

(12) Repeat steps 3 through 11 using different j_T guesses until sufficiently close bracketing guesses have been found. Two guesses bracket the actual value if they produce δj_h 's having opposite signs. Bracketing guesses are sufficiently close when V and I calculated from the two guesses both agree, within the required precision. It is often necessary for bracketing guesses to have four- or five-digit agreement in order for the two V estimates to have three-digit agree-

ment. When the required agreement has been obtained, plot the I-V point and go back to step 2 for additional points.

5.5 Equation Summary for the p⁺/n Diode

All equations, excluding those listed in Section 5.2 and geometric information that must be supplied by the reader, are listed here for the n-type substrate diode.

Starting with the doping densities N_A (p-side) and N_D (n-side), the low field mobilities $\mu_{o,h}$ and $\mu_{o,e}$, the saturation velocity v , the thermal voltage V_T , the elementary charge q , and the dielectric constant ϵ , other constants are calculated from

$$D_h = V_T \mu_{o,h} , \quad D_e = V_T \mu_{o,e} , \quad D^* = 2D_h D_e / (D_h + D_e) \quad (5.18a)$$

$$\sigma_o = (q/V_T) D_e N_D , \quad A_o = D_e N_D / (D_h + D_e) \quad (5.18b)$$

$$a_h = 1 / (q V_T \mu_{o,h}) , \quad V_T b = 1 / (q v) . \quad (5.18c)$$

Boundary values P_2 and V_2 must be solved. The parameters A and E are defined in terms of P_2 and V_2 by

$$E = [\ln(1 + 2P_2/N_D)]^{-1} \quad \text{if } V_2 = 0 \quad \text{and } P_2 > 0 . \quad (5.19a)$$

If $V_2 \neq 0$ and $V_2 \neq 2V_T P_2 / N_D$, use

$$Z_1 \equiv - (V_T/V_2) (1 + 2P_2/N_D) , \quad Z_2 \equiv - V_T/V_2 \quad (5.19b)$$

$$E = H(Z_1, Z_2) \quad (5.19c)$$

where the special function H is defined in Appendix B. For any

case such that $V_2 \neq 2V_T P_2 / N_D$, use either

$$A = (N_D/2) [1 + (V_2/V_T) E] \quad (5.19d)$$

or

$$A = P_2 e^{-1/E} (1 - e^{-1/E})^{-1} . \quad (5.19e)$$

The two equations for A give the same result in theory, but the second should be used if $(V_2/V_T)E$ is so nearly equal to -1 that the first requires more numerical precision than is available. Otherwise, the first can be used. The functions Ω_u , ϕ , and Ω are defined by

$$\text{div grad } \Omega_u = 0 \text{ in sub. , } \Omega_u = 0 \text{ on electrode, } \Omega_u = 1 \text{ on DRB}$$

$$\text{div grad } \phi = -g/D^* \text{ in sub. , } \phi = 0 \text{ on electrode, } \phi = 0 \text{ on DRB}$$

$$\Omega = [P_2 - (N_D/2V_T) V_2] \Omega_u$$

with reflective boundary conditions on the insulated boundaries tacitly assumed. The electrical resistance between electrode and DRB produced by the uniform conductivity σ_0 is R_0 . The ambipolar diffusion currents G_1 and G_2 are given by

$$G_i = -q D^* \int_{S_i} \text{grad } \phi \cdot ds \quad (i = 1, 2)$$

with the unit normal vector chosen so that G_i is positive. R_0 and the G 's may depend on the DR width W .

An approximation for P applicable when $P \gg N_D/2$ throughout most of the substrate is P^∇ given by

$$P^\nabla = (N_D/2) F(2A/N_D, 2(\Omega+\phi)/N_D)$$

where the special function F is discussed in Appendix C. The simpler generalized ambipolar approximation is useful for visualization when $P_2 \gg N_D/2$. If $V_2 \geq 0$, the approximation is

$$P^\nabla \approx \Omega + \phi \quad \text{if } V_2 \geq 0 \quad \text{and} \quad P_2 \gg N_D/2.$$

If $V_2 < 0$, there is an AR and HRR separated by an ARB, which is the constant $\Omega + \phi$ surface characterized by

$$\Omega + \phi = - (N_D/2V_T) V_2 \quad \text{defines ARB} . \quad (5.20)$$

P^∇ is small in the HRR, but the approximation in the AR is

$$P^\nabla \approx \Omega + \phi + (N_D/2V_T) V_2 \quad \text{in AR if } V_2 < 0 \quad \text{and} \quad P_2 \gg N_D/2 .$$

Approximations for the currents are obtained by first defining Γ^∇ by

$$\text{div} [(P^\nabla + A) \text{grad } \Gamma^\nabla] = 0 \quad \text{in substrate}$$

$$\Gamma^\nabla = 0 \quad \text{on electrode,} \quad \Gamma^\nabla = 1 \quad \text{on DRB} .$$

The currents are approximated by

$$j_e A_D \approx (1 - A/N_D) (2V_T P_2/N_D - V_2)/R_0 - G_2 \\ + 2D_e q [(A_0 - A)/N_D] \int_{\text{sub}} \text{grad } \Gamma^\nabla \cdot \text{grad } \phi \, d^3x$$

$$(D_e/D_h) j_h = j_e + (1 + D_e/D_h) G_2/A_D - (2V_T P_2/N_D - V_2)/(A_D R_0) .$$

Calculations are manageable in three dimensions if g can be expressed as

$$g = \alpha(\Omega_u) \text{ grad } \Omega_u \cdot \text{ grad } \Omega_u \quad (5.21)$$

for some function α , which is used to construct the user-supplied function β and derivative β' given by

$$D^*(N_D/2) \beta(v) = v \int_0^1 \int_0^{v_2} \alpha(v_1) dv_1 dv_2 - \int_0^v \int_0^{v_2} \alpha(v_1) dv_1 dv_2 \quad (5.22a)$$

$$D^*(N_D/2) \beta'(v) = \int_0^1 \int_0^{v_2} \alpha(v_1) dv_1 dv_2 - \int_0^v \alpha(v_1) dv_1 \cdot \quad (5.22b)$$

The modified β is given by

$$\beta_m(v) = \beta(v) + (2P_2/N_D - V_2/V_T) v \quad (5.22c)$$

so that

$$\Omega + \phi = (N_D/2) \beta_m(\Omega_u) \quad (5.23)$$

$$P^\nabla = (N_D/2) F(2A/N_D, \beta_m(\Omega_u)) \cdot \quad (5.24)$$

The currents are now approximated by

$$\begin{aligned}
j_e \approx & [(N_D - A)/N_D] (2V_T P_2/N_D - V_2)/(A_D R_O) \\
& + V_T [(N_D - A_O)/N_D] \beta'(1)/(A_D R_O) \\
& + V_T [(A_O - A)/N_D] (INT2/INT1)/(A_D R_O) \quad (5.25a)
\end{aligned}$$

$$(D_e/D_h) j_h = j_e - [2V_T P_2/N_D + V_T \beta'(1) - V_2]/(A_D R_O) \quad (5.25b)$$

where the two integrals INT1 and INT2 are evaluated by selecting a moderately large M (e.g., 100) and using

$$X_1 = 2A/N_D \quad (5.26a)$$

$$B_i' = \beta'(i/M) \quad i=0, \dots, M \quad (5.26b)$$

$$B_i = \beta(i/M) + (2P_2/N_D - V_2/V_T) (i/M) \quad i=0, \dots, M \quad (5.26c)$$

$$C_i = [(N_D/2) F(X_1, B_i) + A]^{-1} \quad i=0, \dots, M \quad (5.26d)$$

$$INT1 \approx (C_0 + C_M)/(2M) + (1/M) \sum_{i=1}^{M-1} C_i \quad (5.26e)$$

$$INT2 \approx (C_0 B_0' + C_M B_M')/(2M) + (1/M) \sum_{i=1}^{M-1} C_i B_i' \quad (5.26f)$$

Note that (5.25b) can be rewritten as

$$V_2 = -A_D R_O (j_e - D_e j_h/D_h) + 2V_T P_2/N_D + V_T \beta'(1) \quad (5.27)$$

One of the DR equations is

$$j_e = q W g_D \quad (5.28)$$

with W the DR width and g_D the value of g at the DR location. Another DR equation is

$$\begin{aligned} \exp(-V_{DR}/V_T) &= N_A^{-1} [P_2 + p_0 - V_T b j_T] \\ &\quad - N_A^{-1} (V_T \epsilon/q) (a_h j_T)^2 [N_D \\ &\quad + V_T b j_T]^{-1} [N_D - V_T b j_T + 2P_2 + 2p_0]^{-1} \quad \text{if } j_T > 0 \end{aligned}$$

$$\exp(-V_{DR}/V_T) = N_A^{-1} (P_2 + p_0) \quad \text{if } j_T \leq 0$$

which can be solved for P_2 in terms of j_T and the DR voltage drop V_{DR} using

$$P_2 = N_A \exp(-V_{DR}/V_T) - p_0 \quad \text{if } j_T \leq 0 \quad (5.29a)$$

$$T_1 \equiv N_A \exp(-V_{DR}/V_T) + V_T b j_T \quad (5.29b)$$

$$T_2 \equiv (V_T/2) (\epsilon/q) (a_h j_T)^2 / (N_D + V_T b j_T) \quad (5.29c)$$

$$T_3 \equiv N_A \exp(-V_{DR}/V_T) + (1/2) V_T b j_T + N_D/2 \quad (5.29d)$$

$$T_4 \equiv T_3 + [T_3^2 + 4T_2]^{1/2} \quad (5.29e)$$

$$P_2 = T_1 - p_0 + 2T_2/T_4 \quad \text{if } j_T > 0 \quad (5.29f)$$

The DR equation used to solve for W is

$$\begin{aligned} W &= (2\epsilon/q)^{1/2} V_{DR}^{1/2} [(N_D + V_T b j_T)^{5/6} \\ &\quad + (2\epsilon/q)^{1/\sqrt{6}} (V_T a_h j_T)^{\sqrt{6}/3} V_{DR}^{-1/\sqrt{6}}]^{-1/\sqrt{6}} \quad \text{if } j_T > 0 \quad (5.30a) \end{aligned}$$

$$W = [(2\epsilon/q) V_{DR}/N_D]^{1/2} \quad \text{if } j_T \leq 0 \quad (5.30b)$$

5.6 Algorithm for Constructing the p^+/n Diode I-V Curve

A suggested algorithm for constructing I-V curves for the n-type substrate diode is listed below. The voltage polarity and direction of current when positive are shown in Figure 5.1. An example of an I-V curve is seen by looking ahead to Figure 6.8 in Chapter 6. The diode delivers power (solar cell operation) when V is positive (a forward-biasing polarity) with I positive (a reverse current produced by photogeneration). The "at $3.0 \mu\text{m}$ " curve for the particular example shown in the figure saturates for V less than about -0.3 volts. Numerical problems will result if we try to extend the curve too far into saturation, because A calculated from (5.19) becomes so close to zero that finite numerical precision fails to distinguish it from zero. But there is no need to extend the plot beyond the point where such a problem first occurs, because such a point is far into saturation. In the opposite extreme of large V, the curve is very steep. Attempting to extend the curve too far in this direction also produces numerical problems because some calculated quantities become extremely sensitive to tiny errors (smaller than machine precision) in other quantities. But there is no need to extend the plot beyond the point where such problems begin to occur, because the current is large enough (in absolute value) to destroy the device. The objective is to plot points in the "range of interest," which is the range that avoids numerical problems and should also be the range that is physically interesting. A suggested algorithm is the following:

- (1) Assign values to q , ϵ/q , V_T , N_A , N_D , $p_0 (=n_i^2/N_D)$, V_C (using (5.2)), R_C , A_D , g_D , and the constants on the left sides of (5.18).
- (2) Select a positive value for V_{DR} . Each selected value will produce one point on the I-V curve. Trial and error is the simplest way to find a V_{DR} value that produces a point in the range of interest. After several I-V points have been plotted, they can guide later selections of V_{DR} values.
- (3) Guess at a value for j_T .
- (4) Use (5.29) to solve for P_2 . Change the value to zero if the presence of p_0 in (5.29) produces a negative value.

(5) Use (5.30) to solve for W and (5.28) to solve for j_e . Then calculate j_h from $j_T - j_e$.

(6) With a value assigned to the DR width W , the substrate geometry is also specified. Assign a value to R_0 . Find a fitting function α that (approximately) satisfies (5.21), and use (5.22) to construct the functions β and β' .

(7) Use (5.27) to solve for V_2 .

(8) Use (5.19) to solve for E . The function subprogram in Appendix B can be appended to any FORTRAN driver code for numerical evaluation of the function H . Note that the computer version of H contains a redundant argument Z_3 for improved numerical accuracy. Before calculating E , first calculate Z_3 from $Z_3 = -2(V_T/V_2)(P_2/N_D)$. Then calculate E from $E = H(Z_1, Z_2, Z_3)$.

(9) Use (5.19) to solve for A . If A is found to be negative, the j_T guess was probably too large. Try a less positive or a more negative j_T . If A is positive but so close to zero that the available numerical precision cannot distinguish it from zero when (5.19e) is used, it is probable that either the j_T guess was too small, or the V_{DR} selection places the I-V point too far into saturation. First try a larger j_T . If convergence (step 12 below) cannot be obtained with j_T large enough to avoid this problem, use a smaller V_{DR} .

(10) Use (5.26) to calculate the integrals $INT1$ and $INT2$. The function subprogram in Appendix C can be appended to any FORTRAN driver code for numerical evaluation of the function F .

(11) Use (5.25a) to calculate a new value for j_e , denoted $j_{e,new}$. Then calculate $\delta j_e \equiv j_{e,new} - j_e$. Calculate I from $j_T A_D$ and then use (5.3) to calculate V .

(12) Repeat steps 3 through 11 using different j_T guesses until sufficiently close bracketing guesses have been found. Two guesses bracket the actual value if they produce δj_e 's having opposite signs. Bracketing guesses are sufficiently close when V and I calculated from the two guesses both agree, within the required precision. It is often necessary

for bracketing guesses to have four- or five-digit agreement in order for the two V estimates to have three-digit agreement. When the required agreement has been obtained, plot the I-V point and go back to step 2 for additional points.

5.7 A Necessary Condition for Saturation

"Saturation" is defined here to mean that the diode current is virtually the same as the total rate that charge is liberated in the device via photogeneration. Looking ahead to Figures 6.3 and 6.8 in Sections 6.2 and 6.3, we see that some I-V curves display saturation while others do not. Now that the DR and substrate equations have been listed together, we can derive a very simple necessary (but not sufficient) condition for saturation. Saturation, strong funneling, a wide HRR, and DR collapse occur together, so the condition derived below can also be regarded as a necessary condition to collapse a DR.

We start with the n^+/p diode where saturation means

$$A_D j_e \approx q \int_{\text{sub}} g d^3x = G_1 + G_2 \quad (5.31)$$

where we have used (2.22b). Using (5.31) and the DR equation (5.15) with the substrate equation (5.7) gives

$$G_2 + q A_D W g_D \approx (D_h/D_e) G_1 + (V_2 + 2V_T P_2/N_A)/R_0$$

which is a necessary and sufficient condition for saturation, but contains unknown boundary values. The only additional information regarding the DR needed to obtain a simpler necessary condition is the fact that the quantity

$$V_2 + 2V_T P_2/N_A$$

is positive. This quantity is obviously positive if V_2 is positive. If V_2 is negative, we have forward-biasing conditions and P_2/N_A will be much larger than $-V_2/V_T$. We may therefore assume that the quantity is positive and the necessary condition becomes

$$G_2 + q A_D W g_D > (D_h/D_e)G_1 \quad (\text{necessary to saturate } n^+/p). \quad (5.32a)$$

The left side of (5.32a) is the rate carriers are generated in the DR plus the rate that carriers flow into the DR as predicted by the ambipolar diffusion equation with homogeneous boundary conditions. On the right side, G_1 is the rate carriers flow to the electrode as predicted by the same equation. The necessary condition states that the rate carriers are generated in the DR or flow into the DR must exceed a certain multiple of the rate they flow to the electrode, as predicted by ambipolar diffusion. This is a statement regarding the spatial distribution of photogeneration and says nothing about the strength of the photogeneration. The condition is satisfied if carrier generation is confined to locations sufficiently close to the MJ. This is clearly not a sufficient condition because it can be satisfied under LILC. But if the condition is not satisfied, the DR will not collapse even if the generation rate is great enough to result in $P_2 \gg N_A$, implying that the latter condition is not sufficient to collapse a DR. This assertion is supported by computer simulation results discussed in Section 6.3.

The analog of (5.32a) for the p^+/n diode is

$$G_2 + q A_D W g_D > (D_e/D_h)G_1 \quad (\text{necessary to saturate } p^+/n). \quad (5.32b)$$

Because $D_e > D_h$, (5.32b) is more difficult to satisfy than (5.32a). DR collapse requires carrier generation to be closer to the MJ for the p^+/n device than required for the n^+/p device. This is our first indication that funneling is more difficult to induce in the p^+/n device. But (5.32a) and (5.32b) are only necessary (not sufficient) conditions and we cannot yet rigorously conclude that the p^+/n device is less susceptible to funneling, although it is, as will be seen in Chapter 6.

6. NUMERICAL EXAMPLES AND CONCLUSIONS

6.1 Introduction

This chapter presents numerical examples to illustrate concepts already discussed and to inspire additional discussion. Unnecessary complexity does not help here, and the examples will be simple. Sections 6.2 and 6.3 treat one-dimensional n^+/p and p^+/n diodes. Section 6.4 treats a simple three-dimensional problem having rotational symmetry. Conclusions are summarized in Section 6.5. Qualitative sketches in Figure 5.1 (Chapter 5) show the polarity convention and the direction of the current when positive. The n^+/p diode delivers power (solar cell operation) when V is negative (a forward-biasing polarity) with I positive (a reverse current produced by photogeneration). The p^+/n diode delivers power when V is positive (a forward-biasing polarity) with I positive (a reverse current). Readers that are not interested in mathematical theory can ignore the paragraphs in the sections below that discuss β and β' .

Comparisons are made between theoretical (or model) predictions and predictions from a computer simulation code called PISCES [5]. Material constants used for the calculations are either default values used by PISCES or are derived from such values. All examples below used the following data (see Sections 5.2, 5.3, and 5.5 for notation):

$$\text{doping density (substrate side)} = 8 \times 10^{14}/\text{cm}^3$$

$$\text{doping density (other side)} = 10^{20}/\text{cm}^3$$

$$R_C = 0$$

$$n_i = 1.5 \times 10^{10}/\text{cm}^3$$

$$V_T = 0.016 \text{ V}$$

$$q = 1.6 \times 10^{-19} \text{ C}, \quad \epsilon/q = 6.536 \times 10^6/\text{V-cm}$$

$$D_h = 13/0 \text{ cm}^2/\text{s}, \quad D_e = 26.0 \text{ cm}^2/\text{s}$$

$$a_h = 4.84 \times 10^{17}/\text{A-cm}^2, \quad a_e = 2.42 \times 10^{17}/\text{A-cm}^2$$

$$V_T b = 3.7 \times 10^{11}/\text{A-cm}$$

PISCES includes a variety of second-order effects, such as band-gap narrowing, several types of recombination mechanisms, and mobilities that depend on a variety of things. Good agreement between model and PISCES predictions indicates that the second-order effects are not important to the quantities of interest in the particular example considered.

6.2 The One-Dimensional n^+/p Diode

We start with the one-dimensional n^+/p diode. Let L be the distance between the electrode and MJ, so $L-W$ is the distance between electrode and DRB, where W is the DR width. Two types of generation rate functions are considered. One is uniform below the MJ, i.e., $g=g_0$ where g_0 is a constant. The total rate per device area that carriers are generated below the MJ for this case is g_0L . For the other case, all carrier generation is confined to a horizontal plane a specified distance x_0 above the electrode, so $g=g_0L\delta(x-x_0)$ where δ is the Dirac delta function, x is the distance from the electrode, and g_0L is the total rate per device area that carriers are generated below the MJ.

The only quantities used in the model that depend on geometry and/or carrier generation are A_D , g_D , R_0 , and the functions β and β' . The DRB area A_D is also the device area and is set equal to 1cm^2 , so that the device current in amps is also the current density in amps/cm^2 . For all cases, we use

$$R_0 = (L - W) / (A_D \sigma_0) .$$

For the uniform case, we have $g_D=g_0$ and

$$\beta(v) = [(L - W)^2 / (N_A D^*)] g_0 (1 - v) v$$

$$\beta'(v) = [(L - W)^2 / (N_A D^*)] g_0 (1 - 2v) .$$

For the delta function case (with generation below the DRB), we have $g_D=0$ and

$$B(v) = [2L/(N_A D^*)] g_0 (L - W - x_0) v \quad \text{if } v < x_0/(L - W)$$

$$B(v) = [2L/(N_A D^*)] g_0 x_0 (1 - v) \quad \text{if } v > x_0/(L - W)$$

$$B'(v) = [2L/(N_A D^*)] g_0 (L - W - x_0) \quad \text{if } v < x_0/(L - W)$$

$$B'(v) = - [2L/(N_A D^*)] g_0 x_0 \quad \text{if } v > x_0/(L - W) .$$

The above information supplements step 6 of the algorithm in Section 5.4. All other steps are explicit and require no additional explanation.

The dimension L is arbitrarily set equal to $5 \mu\text{m}$ in the examples below. (It could be made larger but must be less than a diffusion length, because recombination is neglected in the substrate.) Examples are only interesting if they show significant deviations from classical theory predictions (implying high-injection-level-conditions), and the generation rate was chosen to be large enough to make this happen. For this particular diode, a uniform generation rate of $g=g_0=1.25 \times 10^{25}/\text{cm}^3\text{-sec}$ suffices. Including the factor of q , the total charge generation rate per device area below the MJ is $1000 \text{ amps}/\text{cm}^2$, which is the device current when saturated.

Figure 6.1 compares model, PISCES, and classical predictions of the I-V curve produced by a uniform generation rate of $1.25 \times 10^{25}/\text{cm}^3\text{-sec}$, and shows that the classical prediction is not very good for this case. The classical prediction uses the classical law of the junction, which is (5.16a) but used for all j_T and with V_{DR} set equal to $V+V_C$. The classical estimate of W is used in (5.15) and to determine the electrode to MJ distance. The classical estimate is (5.17b) but used for all j_T and with V_{DR} set equal to $V+V_C$. The minority carrier substrate current is calculated by neglecting the drift term and calculating the carrier density from the minority carrier diffusion equation. It could be argued that classical theory is not being given a fair chance, because the ambipolar diffusion equation may be more

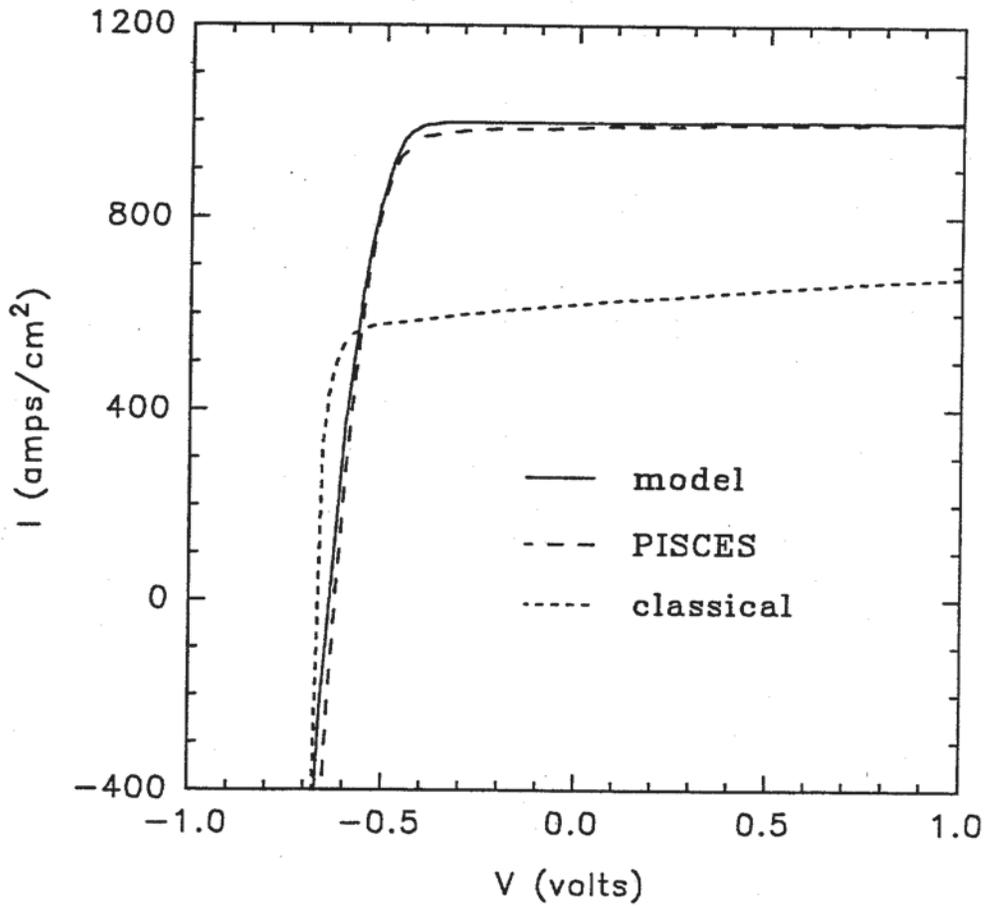


Figure 6.1: Comparison of I-V curve predictions for the n^+/p diode with a uniform $g = 1.25 \times 10^{25} \text{ cm}^{-3} \text{ s}^{-1}$.

appropriate than the minority carrier diffusion equation for calculating carrier density. It turns out that the agreement in Figure 6.1 would be improved if the ambipolar diffusion equation was used for this calculation. But this improvement is not very satisfying in view of the fact that the very same "fix" will worsen the agreement for the p⁺/n diode under high-injection-level conditions treated in the next section. The best agreement obtainable from classical theory for the latter case is produced by the minority carrier diffusion equation. For consistency, this equation is used for all classical theory predictions.

The model and PISCES predictions in Figure 6.1 show that saturation ($I \approx 1000$ amps/cm²) is reached even at some negative voltages. Saturation is an indication that funneling is very strong, but a better indication is obtained by looking at conditions (carrier density and voltage drops) inside of the device. The I-V point at V=1 volt is characterized by the following model-predicted parameters:

$$\text{DR width (W)} = 0.384 \text{ } \mu\text{m}$$

$$\text{substrate voltage drop (V}_2\text{)} = 1.627 \text{ volts}$$

$$\text{electron density at DRB (P}_2\text{)} = 9.171 \times 10^{15}/\text{cm}^3$$

$$A = 6.123 \times 10^{-12}/\text{cm}^3$$

A model-predicted estimate of electron density is P^∇ (given by (5.11)), which is plotted from the above data against distance from MJ in Figure 6.2. The PISCES prediction is also shown. The PISCES prediction places the DRB closer to the MJ than the model prediction. (The DRB and ARB locations shown in the figure are model predictions.) This is consistent with the fact that PISCES calculates a smaller DR voltage drop (V_{DR}) than the model, and is probably due to band-gap narrowing, which PISCES includes but the model does not. Fortunately, this does not seem to affect the I-V curve in Figure 6.1. A compensating correction in the equilibrium built-in potential V_C allows PISCES and the model to agree on the device voltage drop V and the substrate voltage drop V_2 , even when they disagree on V_{DR} . If we account for the shift in DRB location, the two curves in Figure 6.2 will agree very well.

A wide HRR is clearly shown in Figure 6.2, implying strong

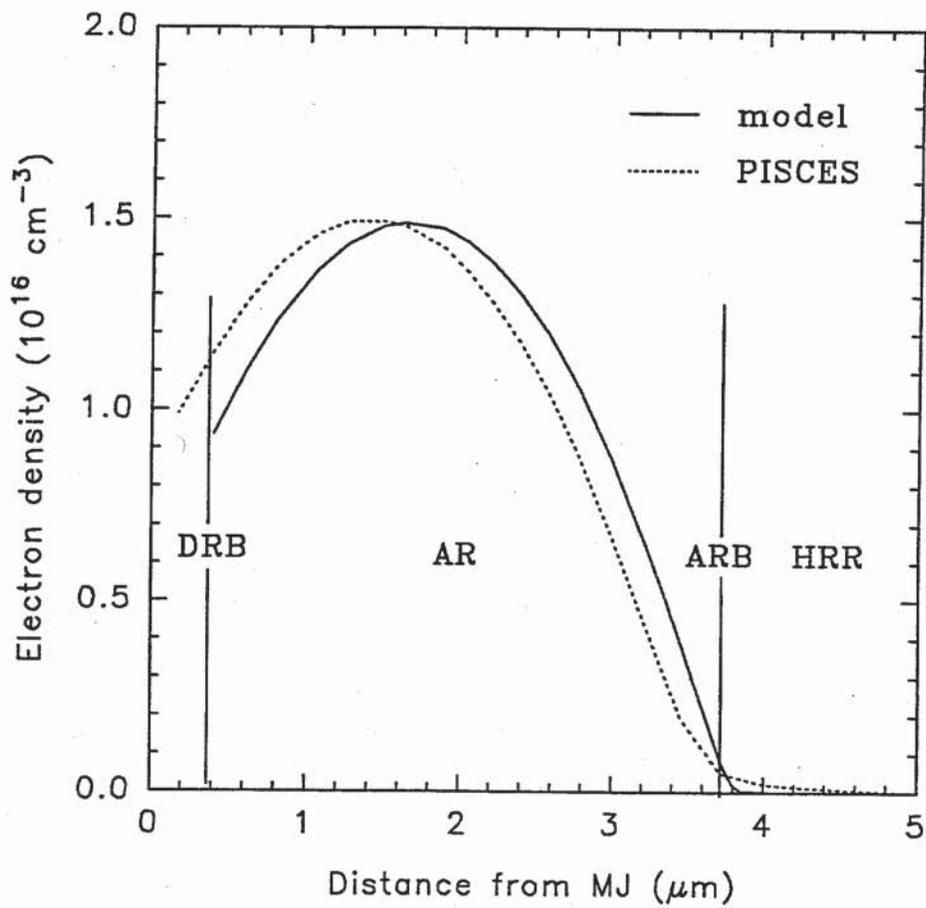


Figure 6.2: Comparison of electron density predictions for the n^+/p diode with a uniform $g = 1.25 \times 10^{25} \text{ cm}^{-3} \text{ s}^{-1}$.

funneling. The V_2 for this configuration is 1.62 (PISCES) or 1.63 (model) volts, which also implies strong funneling. About two-tenths of a volt is across the AR, with the remainder across the HRR, consistent with the statement that most substrate resistance is in the HRR.

The effect of carrier generation location is interesting. The Figure 6.1 model curve is replotted in Figure 6.3, together with an I-V curve produced when all carrier generation is confined to a horizontal plane $2.5 \mu\text{m}$ above the electrode (more than $1 \mu\text{m}$ below the unperturbed DRB for biasing voltages up to 0.5 volts). The total generation rate below the MJ is the same for both cases. The I-V curves are so nearly identical that they could not be distinguished if smooth curves were drawn. Discrete points are shown to emphasize that there really are two data sets here, they just happen to lie on the same curve. It should not be concluded that the model predicts the uniform and "at $2.5 \mu\text{m}$ " cases to be equivalent. We can see differences if we look inside of the device, e.g., the DR and substrate voltage drops are individually different even when they have the same sums. Furthermore, classical theory predicts a slightly larger current for the "at $2.5 \mu\text{m}$ " case. Therefore, there should be some difference between the two curves, but the difference is too small to be seen in the figure.

Saturation in the "at $2.5 \mu\text{m}$ " curve implies that strong funneling is induced at a distance, i.e., by carriers generated outside of the DR. To get the funneling process started, carriers must first diffuse to the DR. Once there, the DR partially collapses and a substrate electric field is created. This field drives more minority carriers to the DR and the funneling process becomes selfsustaining. Figure 6.3 also shows the case where all carrier generation is $1 \mu\text{m}$ above the electrode. Classical theory predicts a comparatively weak current for this case, because most carriers diffuse to the electrode where they recombine. The model shows that funneling is now diminished and no longer strong enough to produce saturation, but still strong enough for the current to be much larger than predicted by classical theory.

Before ending this section, it should be verified that the model, PISCES, and classical predictions all come together under low-injection-level conditions. Such conditions are produced in the diode considered here by decreasing the carrier generation rate by two orders of magnitude. Figure 6.4 compares the predictions for the uniform but reduced generation rate and verifies

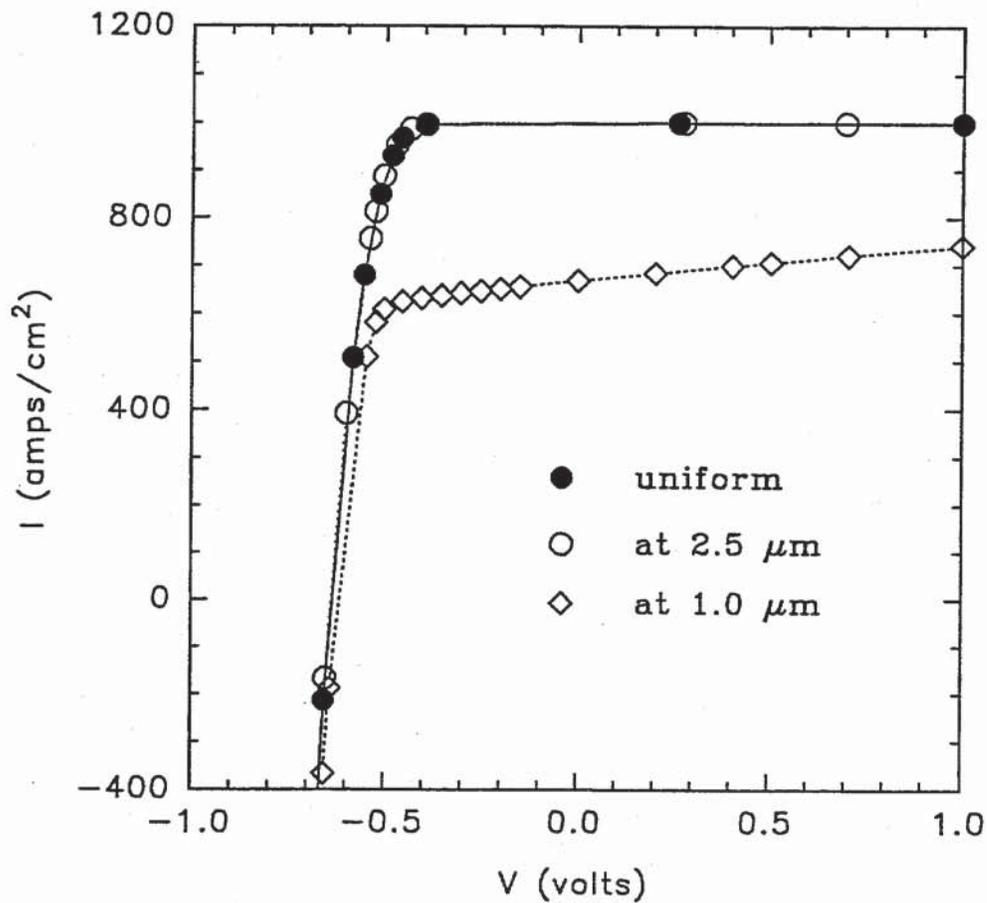


Figure 6.3: Comparison of model-predicted I-V curves for the n^+/p diode when carrier generation location is varied. One curve is produced by a uniform $g = 1.25 \times 10^{25} \text{ cm}^{-3} \text{ s}^{-1}$ (same as Fig.6.1). For the other two curves, all carriers are generated at the indicated distance above the electrode. The total generation rate below the MJ is the same for all cases.

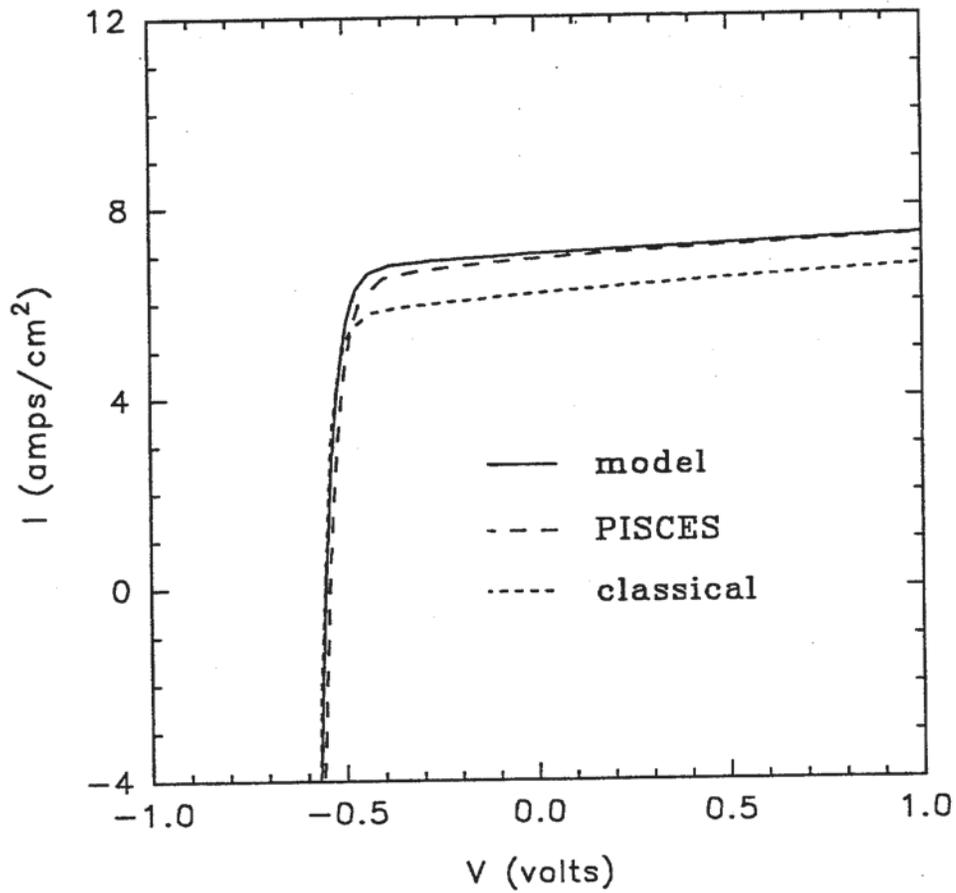


Figure 6.4: Comparison of I-V curve predictions for the n⁺/p diode with a reduced uniform $g = 1.25 \times 10^{23} \text{ cm}^{-3} \text{ s}^{-1}$.

that the predictions do come together.

6.3 The One-Dimensional p^+/n Diode

We now consider the same problem treated in the last section, except that n-type and p-type are interchanged. Figure 6.5 compares model, PISCES, and classical predictions of the I-V curve produced by a uniform generation rate of $1.25 \times 10^{25}/\text{cm}^3\text{-sec}$. The most noticeable difference between Figures 6.5 and 6.1 is that the p^+/n diode is not saturating and the classical prediction is fairly good (although the classical prediction would not be as good if the ambipolar diffusion equation replaced the minority carrier diffusion equation, as discussed in the last section). Compared to the n^+/p diode under the same conditions, funneling is greatly reduced for the p^+/n diode.

A closer comparison can be seen if the n^+/p and p^+/n curves are plotted on the same axis by replacing V with the bias voltage V_B , where $V_B = V$ for the n^+/p diode and $V_B = -V$ for the p^+/n diode. In either case, reverse currents are positive and a positive V_B is a reverse-biasing polarity. The plot is shown in Figure 6.6.

Classical theory predicts the p^+/n device to have the larger (more positive or less negative) current at small V_B , with the curves coming together at larger V_B . This is understandable because the classical current is the sum of a forward current associated with biasing and a reverse current associated with photogeneration. The minority carrier currents, associated with photogeneration, at the electrode and DRB add up to the total generation rate in the substrate; the way this rate is divided between the currents at the two locations depends upon the spatial distribution of photogeneration, but not on mobility. The reverse current associated with photogeneration does not depend on mobility (mobility divides out of the equations). But the forward current is reduced by a reduced minority carrier mobility, so the device having the smaller minority carrier mobility (the p^+/n diode) will have the larger net reverse current, unless the forward currents are negligible so that the two devices have the same currents. This is the classical prediction.

The model prediction in Figure 6.6 agrees with the classical prediction in that the p^+/n device has the larger current at

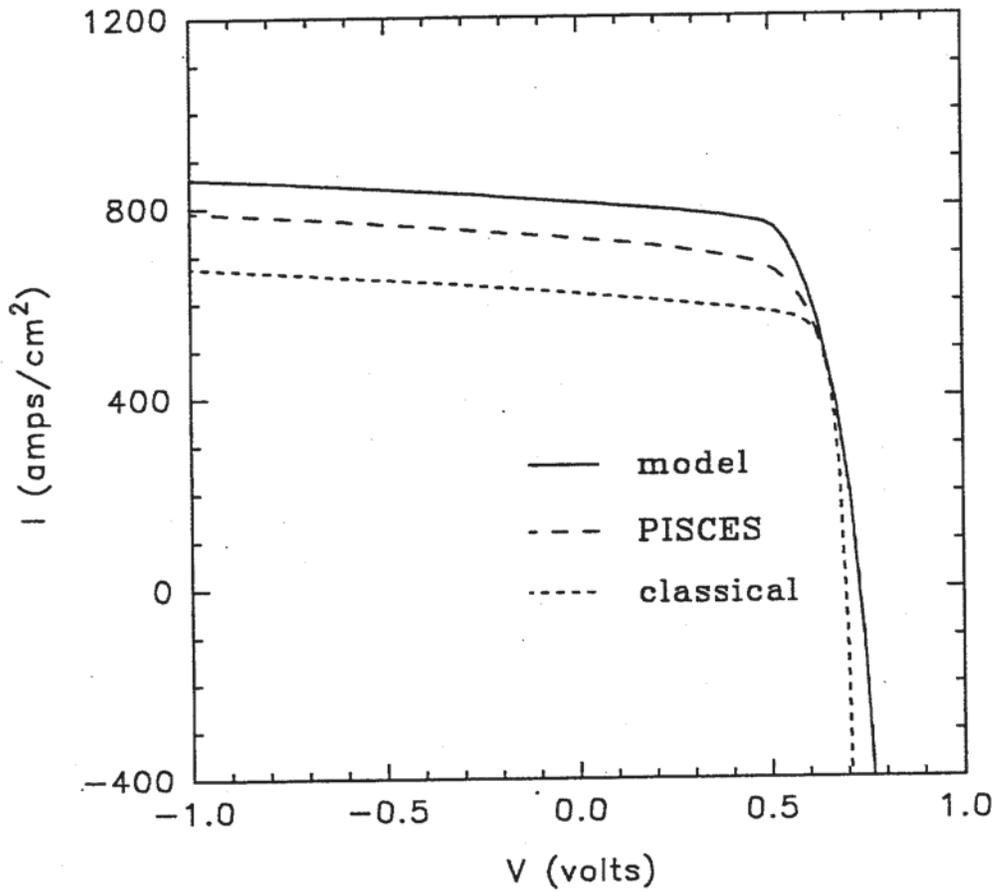


Figure 6.5: Comparison of I-V curve predictions for the p⁺/n diode with a uniform $g = 1.25 \times 10^{25} \text{ cm}^{-3} \text{ s}^{-1}$.

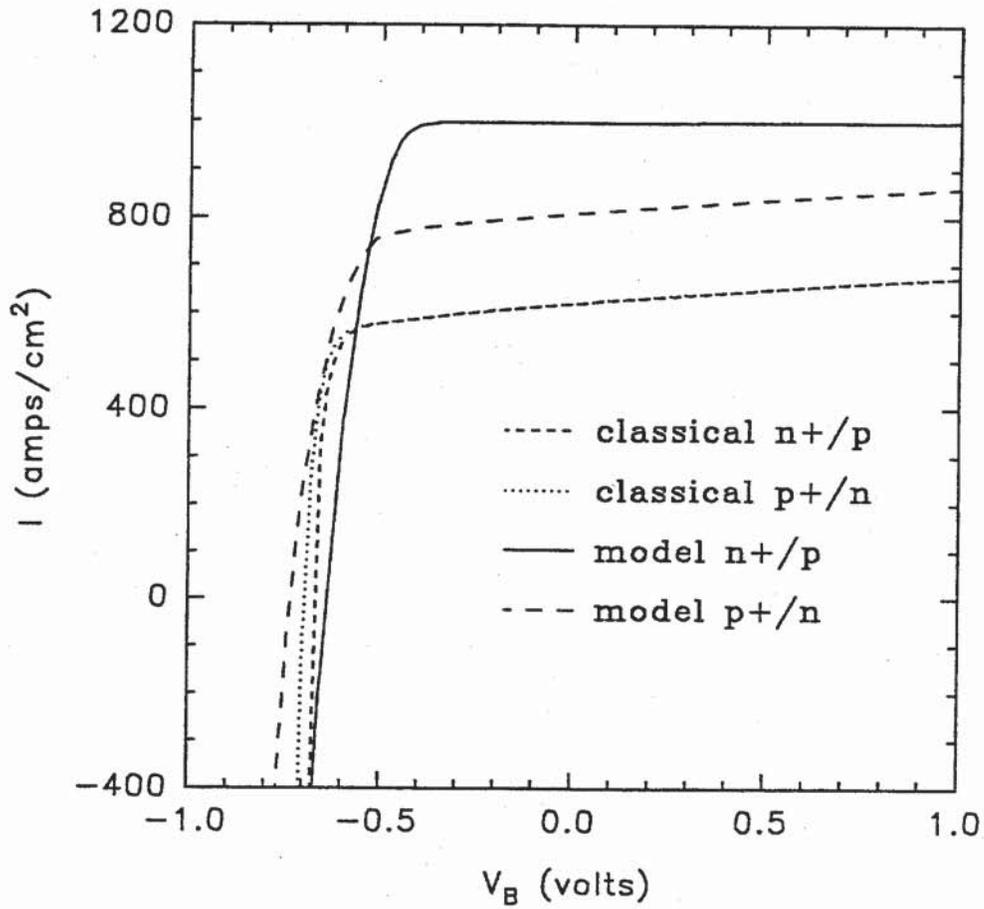


Figure 6.6: Comparison of n^+/p and p^+/n diode I-V curves with a uniform $g = 1.25 \times 10^{25} \text{ cm}^{-3} \text{ s}^{-1}$.

small V_B . But for larger V_B , funneling becomes stronger in the n^+/p device and now this diode has the larger current. Depending on bias voltage and whether carrier generation is sufficient to produce funneling in at least one device, either device can have the larger current.

When looking at either I-V points or I-V curves associated with different conditions, we may see a gradual transition between nonsaturation and saturation, and the two cases may not look so profoundly different. The two model points at $V_B=1$ volt in Figure 6.6 are not really very different. A more profound difference is seen if we look inside of the device at the carrier density and voltage drops. The p^+/n point is characterized by the following model predicted parameters:

$$\text{DR width (W)} = 1.123 \text{ } \mu\text{m}$$

$$\text{substrate voltage drop (V}_2\text{)} = -0.108 \text{ volts}$$

$$\text{hole density at DRB (P}_2\text{)} = 3.682 \times 10^{15}/\text{cm}^3$$

$$A = 3.760 \times 10^{13}/\text{cm}^3$$

The above data were used to plot the hole density in Figure 6.7, which also shows the PISCES prediction. PISCES predicts V_2 to be -0.113 volts, which is nearly the same as the model prediction. The agreement between the model and PISCES predictions looks good in Figure 6.7.

Comparing Figure 6.7 and a V_2 value of about -0.11 volts to Figure 6.2 and a V_2 value of about 1.63 volts, we can now see striking differences between the two cases. The DR is collapsed and the substrate voltage drop is large for the n^+/p case. But for the p^+/n case, the DR is wide and supports nearly all of the applied plus built-in potential, with only a small fraction of this potential across the substrate. The n^+/p case shows a wide HRR. There is a theoretically predicted HRR for the p^+/n case, but it is so narrow as to be almost nonexistent. Because this HRR is so narrow, the substrate voltage is across a highly conductive region. This high conductivity nearly compensates for the smallness of V_2 , so that funneling is occurring in this nonsaturated p^+/n diode and the current is almost as large as in the saturated

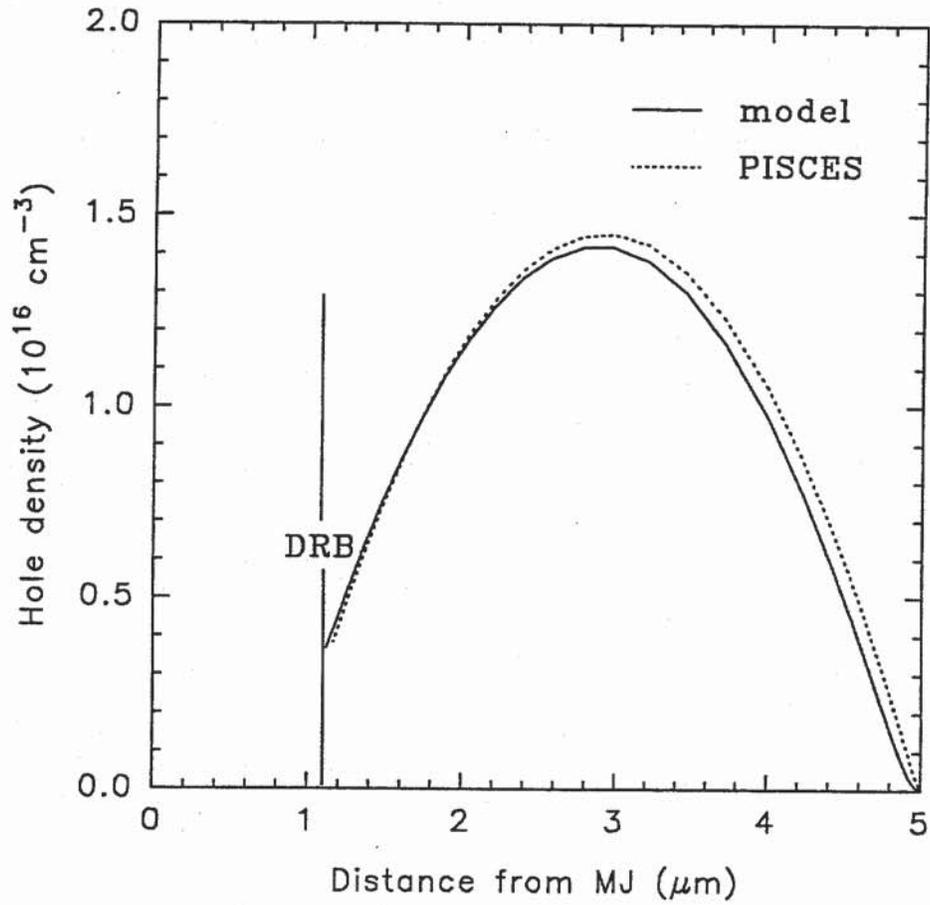


Figure 6.7: Comparison of hole density predictions for the p^+/n diode with a uniform $g = 1.25 \times 10^{25} \text{ cm}^{-3} \text{ s}^{-1}$.

n^+/p diode. If the strength of funneling is measured by the size of the currents, then funneling is not greatly different for the two cases. If the strength of funneling is measured by the size of the substrate voltage drop (which is the convention used when strong funneling is equated to saturation), then funneling is greatly different for the two cases.

With the exception of a region close to the electrode, the minority carrier density in Figure 6.7 greatly exceeds the doping density, even at the DRB. It is interesting (perhaps surprising) that this is not sufficient to collapse the DR. The fact that the DR has not collapsed (enough for the substrate voltage to be great enough to produce saturation) can be predicted from the fact that the necessary condition (5.32b) is not satisfied. The condition can be satisfied if carriers are generated closer to the MJ. If all generation is moved to a horizontal plane $3 \mu\text{m}$ above the electrode, the necessary condition will be satisfied at any point on the I-V curve where the DR width W exceeds $0.5 \mu\text{m}$. Assuming the generation rate is great enough to satisfy all other necessary conditions (whatever they are), we can expect to see saturation somewhere on the I-V curve. This is seen in the "at $3.0 \mu\text{m}$ " curve in Figure 6.8. A close look at this curve finds a small but rapid change in slope at $V \approx -0.3$ volts. It seems reasonable to call this point the onset of saturation. The DR width near this point is between 0.78 and $0.74 \mu\text{m}$ (depending on the exact location of the onset point), so the necessary condition (5.32b) is fairly close to (but not quite) a sufficient condition for this example.

Figure 6.8 also shows the I-V curve produced when all carrier generation is $2.5 \mu\text{m}$ above the electrode. The difference between this and the uniform case is large enough to be visible in the figure, but still very small. The "at $2.5 \mu\text{m}$ " curve does not saturate, even though the generation location is only $0.5 \mu\text{m}$ away from that for the saturating "at $3.0 \mu\text{m}$ " curve. At $V = -1$ volt, the substrate voltage drop for the "at $2.5 \mu\text{m}$ " case is about -0.11 volt (almost the same as the uniform case), compared to -0.44 volt for the "at $3.0 \mu\text{m}$ " case. The currents for the two "at" cases are almost the same. This is another illustration of the fact that the difference between nonsaturation and saturation is more profound if we look at substrate voltage drops instead of currents.

The final noticeable difference between Figures 6.1 and 6.5 is

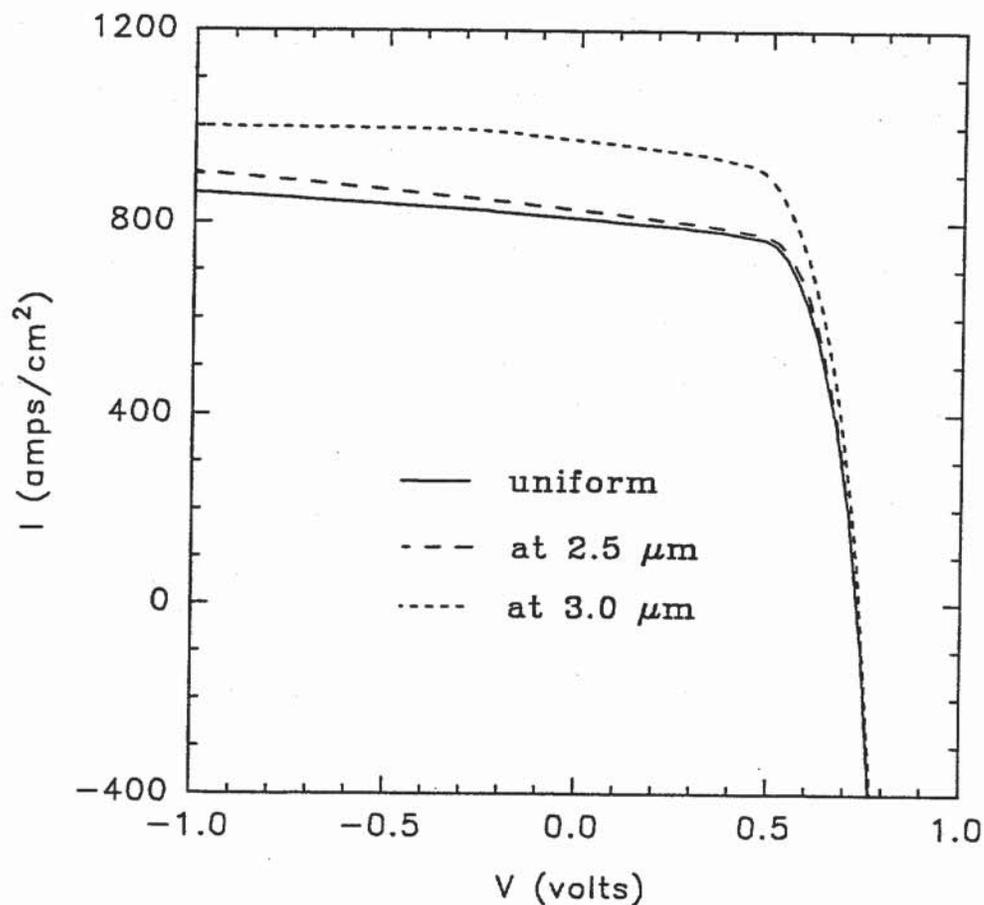


Figure 6.8: Comparison of model-predicted I-V curves for the p^+/n diode when carrier generation location is varied. One curve is produced by a uniform $g = 1.25 \times 10^{25} \text{ cm}^{-3} \text{ s}^{-1}$ (same as Fig.6.5). For the other two curves, all carriers are generated at the indicated distance above the electrode. The total generation rate below the MJ is the same for all cases.

that the model does not agree as well with PISCES in the latter figure. This might be explained in terms of sensitivity. The condition represented in Figure 6.5 is close to some kind of threshold, in the sense that the device is trying to saturate but cannot quite do so. An HRR is wide enough to influence the minority carrier current at the electrode, but not wide enough to either produce saturation or to be clearly visible in Figure 6.7. The calculated minority carrier current at the electrode is sensitive to error in the minority carrier density near the electrode where the density is small. It was argued in Section 4.4 that if $P^{\nabla} \approx P \gg N_D$ almost everywhere, then $P^{\nabla} \approx P$ everywhere, even near the electrode. This is still true, but we must distinguish P governed by the quasi-neutral equations from the PISCES-calculated minority carrier density, which is governed by a more complicated set of equations. While the model- and PISCES-predicted minority carrier densities agree well in terms of absolute error, the relative or fractional error is significant near the electrode. We should expect some error when an HRR strongly influences the minority carrier current but does not block it, i.e., when conditions are almost but not quite able to produce saturation. But even under these adverse conditions, the agreement between the model and PISCES curves in Figure 6.5 is fairly good.

6.4 A Simple Three-Dimensional Diode

A simple three-dimensional example is considered, primarily to illustrate a general method for treating such problems. The objective is to illustrate the method while avoiding difficult integrals, so the example is highly idealized. Readers that are willing to evaluate difficult integrals can apply the method to more difficult problems.

In this example, one DRB is isolated from all other DRBs. The DRB is a circular disk of radius r_D and photogeneration is confined to a circular cylinder having the same radius r_D and length L . The cylinder is normal to the device and centered on the DRB. It is assumed that r_D and L are both small compared to the DRB-to-electrode distance. Because recombination is neglected, r_D and L are both required to be small compared to the diffusion length. As long as the above conditions are satisfied, it is not required that the DRB-to-electrode distance be small compared to

the diffusion length. We can neglect recombination and regard the electrode as infinitely far away, so the problem to be solved reduces to that shown in Figure 6.9, which also shows the coordinate system. The generation rate is uniform and equal to g_0 (a constant) inside of the cylinder. Cases in which the cylinder radius is less than r_D might be approximated by the case considered here if g_0 is selected to produce the same total generation rate per unit length in the vertical direction. The special choice of r_D for the cylinder radius simplifies some integrations. A better representation of a possible physical arrangement would use a generation function that is exponentially attenuated in the vertical coordinate. The attenuated problem is left for any reader that is willing to evaluate the required integrals.

The DR width W is simulated by retaining the flat disk geometry but reducing the generation cylinder length from L to $L-W$ (assuming that $L > W$). A majority carrier current calculated from (5.15) or (5.28) compensates for the missing cylinder section. For notational brevity, a length L is used in the analysis and then replaced with $L-W$ in the final equations. The DR width can also add to the lateral dimension r_D in three dimensions, but this is ignored in the analysis below. No distinction is made between the DRB radius and the MJ radius.

The only quantities, used in the algorithms in Sections 5.4 and 5.6, that depend on geometry and/or carrier generation are A_D , g_D , R_0 , and the functions β and β' . We obviously have $g_D = g_0$ and $A_D = \pi r_D^2$. R_0 is well known for the flat circular disk and given by $R_0 = 1/(4\sigma_0 r_D)$. The analysis is finished when the functions β and β' have been constructed. These functions are derived from α satisfying (5.8). But there is no such α for this three-dimensional problem and fitting is required. The definition of a "best fit" is somewhat arbitrary, but a particular definition will produce exact calculations of the ambipolar diffusion currents G_1 and G_2 . This is demonstrated below for arbitrary geometries and generation functions. Readers that are not interested in mathematical theory can go directly to the paragraph following the equations for β and β' on page 95.

A sufficient condition for a fit to g to produce the correct G 's is found by using (2.19), (2.20), and the divergence theorem to write (2.22a) as

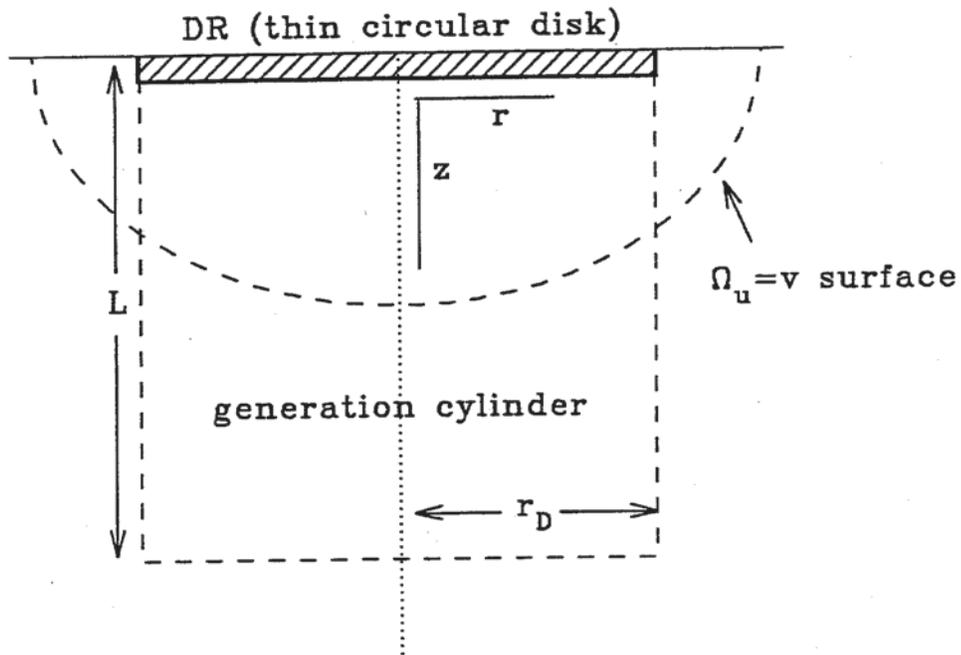


Figure 6.9: A simple three-dimensional geometry. Generation is confined to the cylinder of length L and radius r_D (same as the DR radius). The electrode is at infinity. The $\Omega_u = v$ surface encloses the region $R_C(v)$.

$$\begin{aligned}
-G_2/(q D^*) &= \int_{S_2} \text{grad } \phi \cdot ds = \int_{S_1} \Omega_u \text{ grad } \phi \cdot ds + \int_{S_2} \Omega_u \text{ grad } \phi \cdot ds \\
&- \int_{S_1} \phi \text{ grad } \Omega_u \cdot ds - \int_{S_2} \phi \text{ grad } \Omega_u \cdot ds = \int_{\text{sub}} \Omega_u \text{ div grad } \phi d^3x
\end{aligned}$$

or

$$G_2 = q \int_{\text{sub}} \Omega_u g d^3x .$$

We can select a number M and partition the substrate into subregions $\delta_1R, \delta_2R, \dots, \delta_MR$, where each δ_iR is the region between the $\Omega_u=(i-1)/M$ and the $\Omega_u=i/M$ surfaces. The integral can be written as the sum

$$G_2 = q \sum_{i=1}^M \int_{\delta_iR} \Omega_u g d^3x \approx q \sum_{i=1}^M (i/M) \int_{\delta_iR} g d^3x$$

with the approximation on the far right becoming exact in the large M limit. A fitting function g_{fit} will produce the same G_2 if it satisfies

$$\int_{\delta_iR} g_{\text{fit}} d^3x = \int_{\delta_iR} g d^3x \quad \text{for all } i = 1, \dots, M$$

which is equivalent to

$$\int_{R_C(v)} g_{\text{fit}} d^3x = \int_{R_C(v)} g d^3x \quad \text{for all } v \in (0,1)$$

where $R_C(v)$ is the region above the $\Omega_u=v$ surface (the subscript denotes compliment to distinguish $R_C(v)$ from $R(v)$ in Section 4.5). If the fitting function has the form (5.8), steps similar

to those that produced (4.26) can be used to write the above equation, for the p-type substrate, as

$$\beta'(v) - \beta'(1) = q (1 + D_h/D_e) (R_o/V_T) \int_{R_C(v)} g d^3x . \quad (6.1)$$

The above equation is used to define the best fit for the p-type substrate with arbitrary geometry and arbitrary g . Interchange D_h and D_e for the n-type substrate. Integrating this equation with respect to v solves for β . The integration constant and $\beta'(1)$ are both determined by the two endpoint conditions $\beta(0)=\beta(1)=0$.

For the special case of the circular disk in Figure 6.9, Ω_u is well known and the $\Omega_u=v$ surface is seen in the figure as an ellipse having the equation

$$r^2 \sin^2(\pi v/2) + z^2 \tan^2(\pi v/2) = r_D^2 \quad (\text{equation of } \Omega_u=v \text{ surface}).$$

Omitting the argument $\pi v/2$ from the trigonometric functions for notational brevity, the integral of g can be written as

$$\int_{R_C(v)} g d^3x = 2\pi \int_0^{r_D} \csc \int_0^{[r_D^2 \cot^2 - r^2 \cos^2]^{1/2}} g dz r dr$$

which integrates in z first. To integrate in r first, it is convenient to make the change in variables $w=r^2 \sin^2(\pi v/2)$ and write the integral as

$$\int_{R_C(v)} g d^3x = (\pi/\sin^2) \int_0^{r_D} \cot \int_0^{r_D^2 - z^2 \tan^2} g dw dz .$$

The above equations apply to arbitrary g . Specializing to the case where $g=g_o$ inside the cylinder and $g=0$ outside, the integral becomes

$$(\pi g_0 r_D^2)^{-1} \int_{R_C(v)} g d^3x = L$$

$$\text{if } \cos^2 > [(L^2/2r_D^2)^2 + L^2/r_D^2]^{1/2} - L^2/2r_D^2 \quad (6.2a)$$

$$(\pi g_0 r_D^2)^{-1} \int_{R_C(v)} g d^3x = L/\sin^2 - (2/3) r_D \cos^4/\sin^3 \\ - (1/3) (L^3/r_D^2)/\cos^2$$

$$\text{if } L^2/(L^2+r_D^2) < \cos^2 < [(L^2/2r_D^2)^2 + L^2/r_D^2]^{1/2} - L^2/2r_D^2 \quad (6.2b)$$

$$(\pi g_0 r_D^2)^{-1} \int_{R_C(v)} g d^3x = (2/3) (r_D/\sin) [\cos + \cos^3/(1 + \cos)]$$

$$\text{if } \cos^2 < L^2/(L^2 + r_D^2) \quad (6.2c)$$

Equation (6.2b) applies when v satisfies the condition that the $\Omega_u=v$ surface intersects the cylinder wall and lower end, i.e., only the lower cylinder "corners" (seen as corners in Figure 6.9) are excluded from $R_C(v)$. The corners contain a small amount of carrier generation and there is no need to retain such complexity for such an unimportant v interval. Therefore (6.2a) will be used over the extended interval $\cos^2 > L^2/(L^2+r_D^2)$. This produces a slight discontinuity in the integral of g , equivalent to redistributing the generation so that the generation in the corners is placed on a surface. The total generation within a region that completely contains the cylinder is not affected by this redistribution. Substituting this simplified version of (6.2) into (6.1), integrating to solve for β , and replacing L with $L-W$ produces the final result

$$L_0 \equiv L - W, \quad L_1 \equiv (L_0^2 + r_D^2)^{1/2}, \quad L_2 \equiv L_0/L_1, \quad L_3 \equiv r_D/L_1$$

$$C_0 \equiv \pi q (1 + D_h/D_e) (R_0/V_T) g_0 r_D^2 = 2A_D R_0 (\sigma_0/N_A) (g_0/D^*)$$

$$C_1 \equiv C_0 L_0, \quad C_2 \equiv (2/3) C_0 r_D$$

$$C_3 \equiv (2C_1/\pi) \arccos(L_2) + (3C_2/\pi) \ln(1/L_3) \\ + (3C_2/\pi) \ln(1 + L_2) - (C_2 L_2/\pi) [2 + 1/(1 + L_2)]$$

$$C_V \equiv \cos(\pi v/2), \quad S_V \equiv \sin(\pi v/2)$$

If $0 < C_V < L_2$ then:

$$B'(v) = (C_2/S_V) [C_V + C_V^3/(1 + C_V)] - C_3$$

$$B(v) = (2C_1/\pi) \arccos(L_2) + (3C_2/\pi) \ln(S_V/L_3) \\ + (3C_2/\pi) \ln[1 + (L_2 - C_V)/(1 + C_V)] - C_3 v \\ - (C_2/\pi) [2(1 + C_V) + 1/(1 + L_2)] [(L_2 - C_V)/(1 + C_V)]$$

If $L_2 < C_V < 1$ then:

$$B'(v) = C_1 - C_3$$

$$B(v) = (C_1 - C_3) v$$

which applies to the p-type substrate. For the n-type substrate, interchange D_h with D_e and replace N_A with N_D in the C_0 equation.

The ambipolar diffusion current G_2 is related to $\beta'(1)$ via (4.26b), and the above results relate $\beta'(1)$ to C_3 which contains the term $\ln(1/L_3)$. This term becomes singular in the limit as $L \rightarrow \infty$ (because $L_3 \rightarrow 0$), so the diffusion current has a logarithmic singularity as the cylinder length is increased without bound in this idealized geometry. In a real device, the finite diffusion length and/or device dimensions will limit the current if L is too large.

The meaning of a "wide" HRR is interesting. The minority carrier density is negligibly small in a region that extends to infinity in this idealized geometry. If spatial distance defines HRR width, there will always be an infinitely wide HRR. But if a wide HRR is to be associated with saturation, we must use something other than spatial distance to measure HRR width. Saturation occurs when nearly all carrier-modulated substrate resistance is in the HRR. A wide HRR will be interpreted to mean that the HRR contains nearly all substrate resistance. Because of spreading effects, a region that extends to infinity need not have much resistance. A wide HRR is not easy to recognize from a plot of carrier density versus rectangular coordinates. But it might be recognized from a plot that shows how much Ω_u drop is across the HRR. This is because a significant Ω_u drop implies that the HRR contains a significant fraction of the equilibrium resistance. But if the HRR contains a sizable fraction of the equilibrium resistance, then it will contain nearly all of the carrier-modulated resistance (because of the comparatively small carrier-modulated conductivity inside of the HRR). It is therefore most informative to plot carrier density against the coordinate v , which is related to the spatial coordinates by $v = \Omega_u(\mathbf{x})$. The actual carrier density will not be a function of v alone, but the surface average density on the $\Omega_u = v$ surface can be plotted against v . The model-predicted density given by (5.11) or (5.24) is a function of v alone (constant Ω_u surfaces are constant carrier density surfaces) because the actual g is replaced by a fit having the form (5.8). The model-predicted density can be regarded as a model-predicted surface average density, and can be plotted against v using (5.11) or (5.24) with Ω_u replaced by v .

To illustrate saturation and a wide HRR, we consider the specific problem of the n^+/p diode characterized by

$$r_D = 5 \mu\text{m}$$

$$L = 10 \mu\text{m}$$

$$g_0 = 6.25 \times 10^{24} / \text{cm}^3\text{-s}$$

The value selected for g_0 is convenient for making a comparison with the first one-dimensional example in Section 6.2. The total charge generation rate (including the factor of q) divided by MJ area is the same for both cases and is 1000 amps/cm². If the device current is normalized by dividing by MJ area (which is convenient for making comparisons with Figure 6.1), the normalized current will be 1000 amps/cm² when the device is saturated.

The model-predicted I-V curve for this three-dimensional example is shown in Figure 6.10. A PISCES prediction is also shown. PISCES requires finite geometries and the device simulated by PISCES has a cylindrical substrate with a 50- μm radius and a 50- μm length. The vertical wall is reflective and the lower end is the electrode. The version of PISCES used here will not accept a g that is uniform inside the cylinder and zero outside. A rough approximation of a step function of the radial coordinate is the function $\exp(-r^2/r_D^2)$, which is the radial dependence used in the PISCES simulation. A finite grid spacing results in the PISCES-calculated total generation rate being different than the actual volume integral of g . An adjusted value was assigned to g_0 so that the model and PISCES calculate the same total generation rate.

Most of the difference between the two curves in Figure 6.10 is due to recombination, which is more noticeable in this extended geometry than in the 5- μm one-dimensional geometry represented in Figure 6.1. The lifetime for Shockley-Reed-Hall (SRH) recombination used by PISCES was 1 microsecond. The difference between the two curves in Figure 6.10 is not really very large, but we have seen better agreement in Figure 6.1. To improve the agreement and verify that other model calculations (e.g., the treatment of a three-dimensional geometry) are okay, PISCES was run again for the same problem, but with Auger recombination calculations turned off and the SRH lifetime changed to 10 milliseconds. This virtually eliminates recombination from the PISCES calculations.

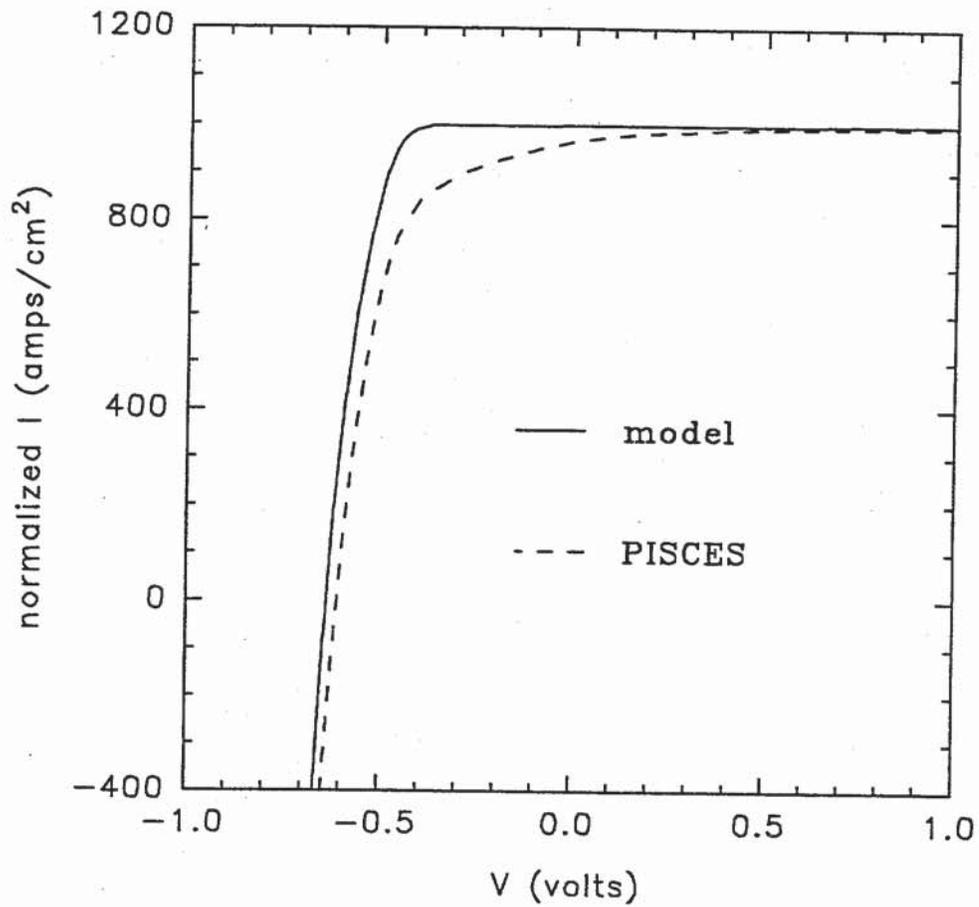


Figure 6.10: Comparison of I-V curve predictions for the three-dimensional n^+/p diode using $r_D = 5 \mu\text{m}$, $L = 10 \mu\text{m}$, and $g_0 = 6.25 \times 10^{24} \text{ cm}^{-3} \text{ s}^{-1}$.

The result is shown in Figure 6.11 and the agreement is now almost as good as it was in Figure 6.1. It is more difficult to control numerical error associated with grid line spacing in three dimensions than in one, and a comparison between PISCES-calculated electron and hole currents at various locations to the generation rate indicates that there are sizable numerical errors in the PISCES curve at small V . Furthermore, the ad hoc fit given to PISCES for the radial dependence of g will not produce the correct G_2 (unlike the fit used by the model which does produce the correct G_2). Both errors are in a direction such that the agreement in Figure 6.11 would be further improved at small V if these errors were eliminated. This indicates that the model is okay except for neglecting recombination.

It is interesting that the model curves in Figures 6.1 and 6.11 are indistinguishable. This is not an accident. The two geometries are almost equivalent as far as the boundary value problems (when recombination is neglected) are concerned. The $5\text{-}\mu\text{m}$ one-dimensional (1D) problem represented by Figure 6.1 can be given the same area πr_D^2 as the three-dimensional (3D) problem by inserting a reflective vertical cylindrical wall. This results in $L=r_D$ for the 1D problem, compared to $L=2r_D$ for the 3D problem. For the 1D problem with uniform g , we have $G_1=G_2$. Calculating the G 's for the 3D problem from $B'(0)$ and $B'(1)$, which are calculated from the equations listed earlier, we find that (when $W \approx 0$) $G_2 \approx 1.06G_1$. The total generation rate is the same for the two problems, so both G 's are nearly the same for the two problems. For the 1D problem, we have $1/R_0 = \pi \sigma_0 r_D$, which is roughly the same as the value $4\sigma_0 r_D$ applicable to the 3D problem. The two geometries are not exactly equivalent, but they are almost equivalent. The I-V curves saturate early (at small V), so they are primarily controlled by total generation rate and are insensitive to other factors such as geometric effects. The small difference in geometry is not observable in these curves.

The I-V point at $V=1$ volt in Figure 6.11 is characterized by the following model-predicted parameters:

$$\text{DR width. (W)} = 0.403 \mu\text{m}$$

$$\text{substrate voltage drop (V}_2\text{)} = 1.607 \text{ volts}$$

$$\text{electron density at DRB (P}_2\text{)} = 5.038 \times 10^{15}/\text{cm}^3$$

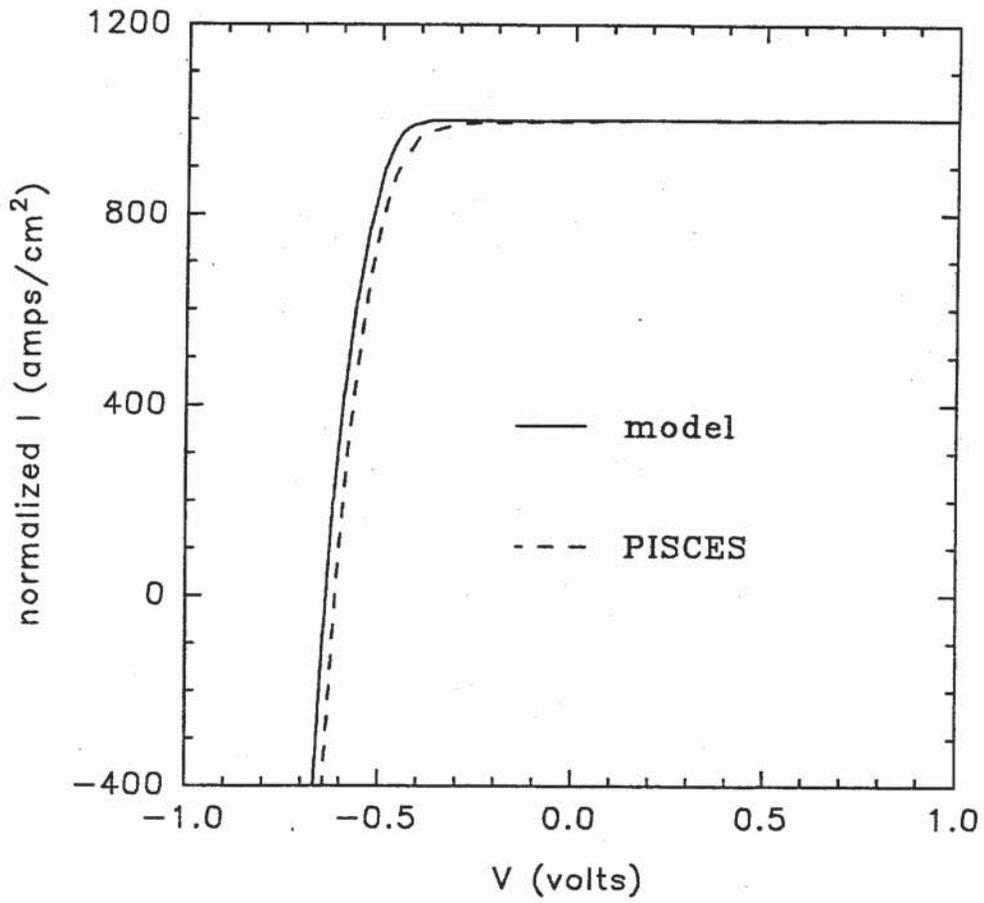


Figure 6.11: Same as Fig.6.10 except that recombination was eliminated from the PISCES calculations.

$$A = 7.256 \times 10^{-12} / \text{cm}^3 .$$

These data were used to plot surface average electron density versus v , as discussed earlier, and the result is shown in Figure 6.12. This curve can be compared to the curve in Figure 6.2 by noting that v is a linear function of distance from MJ in the latter figure, satisfying $v=1$ at the DRB and $v=0$ at the electrode. The two geometries are not exactly equivalent and the two curves are not identical. But plotting density against v instead of spatial distance makes a similarity visible. The two curves show about the same HRR width if "width" is measured by v instead of spatial distance.

6.5 Conclusions

Strong funneling is loosely defined by the condition that the DR has collapsed and there is a large substrate voltage drop. Strong funneling, saturation, and a wide HRR occur together under steady-state conditions. A wide HRR in a three-dimensional geometry may be easiest to recognize if surface average minority carrier density on the $\Omega_u=v$ surface is plotted against v . When strong funneling occurs, most substrate voltage is across the HRR which limits the current. Because of the high resistance in a wide HRR, the current under saturation conditions need not be much larger than under nonsaturation conditions. If some parameter (e.g., the photogeneration rate or a device dimension) is varied, the transition between nonsaturation and saturation can appear very gradual when looking at terminal currents. Although still continuous, the transition appears more abrupt when looking at substrate voltage drop or HRR width, and onset conditions can be reasonably well defined. The presence of a wide HRR implies that the ambipolar diffusion equation fails to provide a good approximation for the carrier density function. This equation might be used in the AR if boundary conditions are modified to account for the presence of the ARB, but a better approximation was provided for quantitative estimates. Another observation is that carriers need not be generated inside of a DR to collapse the DR. Strong funneling can be induced at a distance, i.e., by carriers generated outside of the DR.

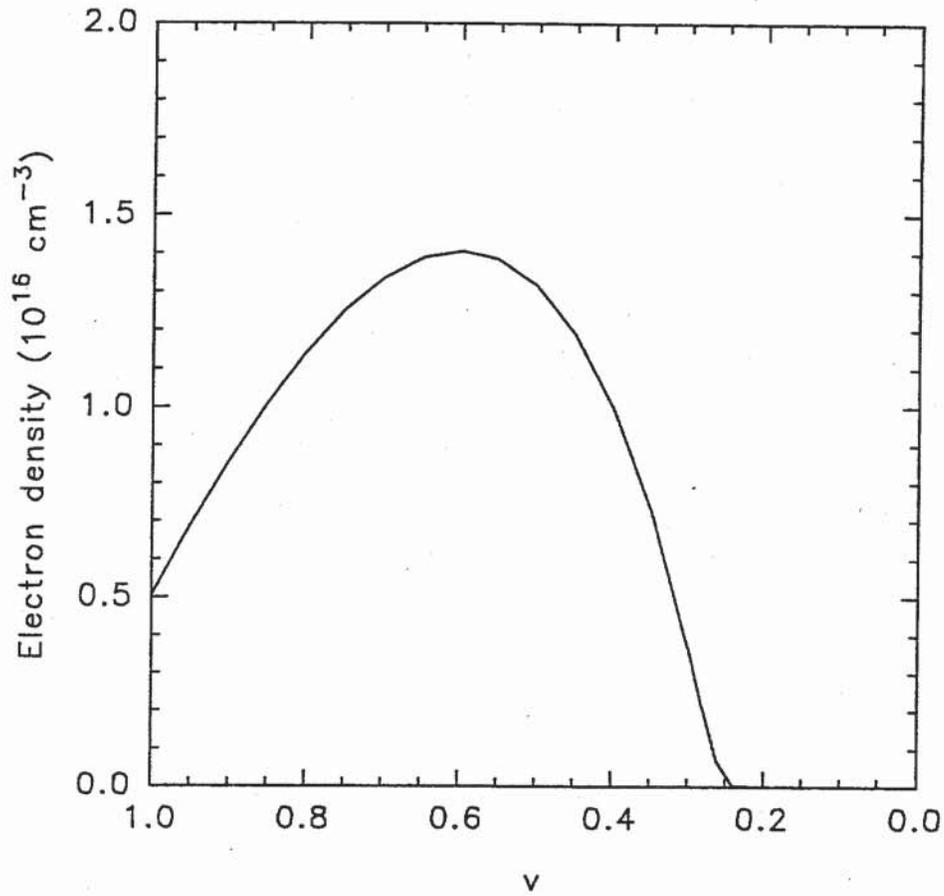


Figure 6.12: Model-predicted surface average electron density on the $\Omega_u = v$ surface, plotted against v for the three-dimensional n^+/p diode.

A necessary condition for saturation (or DR collapse) was derived in terms of ambipolar diffusion currents and is a statement regarding the spatial distribution of photogeneration. The necessary condition is satisfied if all generation is confined to a region sufficiently close to the MJ. The condition is not satisfied, and a wide HRR cannot form, if the generation is too strong at locations too close to the electrode. A saturated condition can be changed to a nonsaturated condition by adding additional generation if the addition is near the electrode. This does not imply that additional generation can decrease the reverse current. The current under nonsaturation conditions can exceed the current under saturation conditions if the former case is produced by a larger generation rate. If the necessary condition for saturation is not satisfied, the DR will not collapse even if the carrier density greatly exceeds the doping density in the substrate and at the DRB. The condition is more difficult to satisfy for the p^+/n diode than for the n^+/p diode. Compared to the n^+/p diode, saturation of the p^+/n diode requires that generation be closer to the MJ. A spatially uniform generation rate will not saturate a one-dimensional p^+/n diode unless the DR width is sufficiently large compared to the substrate thickness. When both diode types are operated under similar conditions and neither saturates, either can have the larger current.

The motivation for this steady-state analysis is to obtain physical and mathematical guidance for a future transient analysis. It may be appropriate to point out some of the similarities between the steady-state and transient problems. The discussion below refers to the transient problem in which funneling is induced by an ion that produces a track of free carriers in the DR and/or substrate.

Strong funneling can still be defined by the condition that the DR has collapsed and there is a large substrate voltage drop. The transient analog of saturation is that the minority carrier current at the electrode is negligible at the time of interest. It is reasonable to expect this condition to accompany strong funneling, and this has been seen in PISCES transient simulation results.

Steady-state funneling can be induced at a distance and there is little distinction between carriers generated within the DR and carriers generated outside of but close to the DR. Transient simulation results show that transient funneling can also be

induced at a distance, i.e., not requiring a direct DR hit. Carriers must first diffuse to the DR to get the funneling process started. Once there, the DR partially collapses and a substrate electric field is created. This field drives more minority carriers to the DR and the funneling process becomes self-sustaining, until the track is sufficiently diminished for the DR to recover. Furthermore, carriers generated within the DR do not have a special significance. We might expect these carriers to be collected much faster than those outside of the DR, because of the strong DR electric field, so that these carriers are distinguishable from the others in terms of charge collection time. But this is not the case. After the ion hit, carriers initially in the DR are separated and driven out as the DR simultaneously completely or partially collapses, but charge collection at the device terminals does not respond fast enough to be significantly affected by this initial charge separation. Terminal currents that contribute most to collected charge are seen while the DR is recovering, with these carriers now outside of the DR and responding to drift and diffusion just like all other nearby carriers. A plot of collected charge (which is the time integral of terminal current and hides current "blips" that negligibly contribute to collected charge) versus time is very smooth and has no demarcation that distinguishes one group of carriers from another. "Prompt charge," discussed in some of the older literature on single event effects, is not well defined.

If the track is not long enough to reach the electrode and strong funneling is occurring, it is reasonable to expect an HRR below the track, with an electric field inhibiting the downward flow of minority carriers. Such an HRR should be depleted of minority carriers and (because of quasi-neutrality) depleted of excess majority carriers. The visual impression is an absence of downward diffusion of the track, and this is seen in transient simulation results.

An interesting case occurs when the track is long enough to reach the electrode. Transient simulations were run for two track lengths in a reverse-biased n^+/p diode. In both cases the DR is circular with a $5\text{-}\mu\text{m}$ radius and was $100\ \mu\text{m}$ above the electrode plane. There was a reflective vertical cylindrical wall with a $50\text{-}\mu\text{m}$ radius. The ion tracks were perpendicular to the device and centered on the DR. Both tracks had the same density but one was $35\ \mu\text{m}$ long while the other reached the electrode. The collected charge versus time curves were almost the same for the two cases.

At very early times (≤ 0.2 ns) after the ion hit, the current produced by the long track was a factor of two to three larger than that for the short track. At later times (≥ 0.6 ns), the currents were nearly the same. The time over which the currents significantly differed was so short that little charge was collected during this time, so the collected charge versus time curves were nearly the same. At 0.6 ns, the currents were nearly the same even though neither track was significantly diminished. This situation is similar to the steady-state situation in which greatly different HRR widths accompany greatly different substrate voltage drops to produce nearly the same currents. For the long-track transient case, an HRR cannot form until the lower end of the track has been cleared away. At 0.6 ns, there has not yet been time for much of the track to be cleared away, and the HRR is narrow. But because of the small substrate resistance, the DR resists collapsing. The DR was only mildly collapsed for the long-track case, compared to a greatly collapsed DR for the short-track case. The narrow HRR seen in the long-track case was accompanied by a correspondingly small substrate voltage drop, which produces nearly the same current as the short-track case. It is interesting that the long track is less able to maintain a collapsed DR than the short track. The spatial distribution condition, necessary for steady-state saturation, may have a transient analog.

APPENDIX A: THE DR EQUATIONS

A1 Introduction

The DR equations derived in Reference [2] are very messy. They are summarized here for the n^+/p and p^+/n junctions, using mostly Reference [2] notation, and then simplified. The following notation is used:

- U_p, U_n = potentials at p-side and n-side DRBs, respectively, relative to potential at metallurgical junction
- n_p, p_p = carrier densities at p-side DRB
- n_n, p_n = carrier densities at n-side DRB
- V_{DR} = potential at n-side DRB relative to potential at p-side DRB
- N_A, N_D = p-side and n-side doping densities, respectively
- j_h, j_e = scalar current densities at the DRB on the lightly doped side; these scalars are positive when currents are directed from the n-side towards the p-side
- W = DR width
- δ_o = unit step function ($\delta_o(x)=0$ if $x<0$ and $\delta_o(x)=1$ if $x\geq 0$)
- $\mu_{o,h}, \mu_{o,e}$ = low field mobilities
- v = saturation velocity
- V_T = thermal voltage
- q = elementary charge
- ϵ = dielectric constant
- $a_e, a_h = 1/(q V_T \mu_{o,e})$ and $1/(q V_T \mu_{o,h})$, respectively
- $V_T b = 1/(q v)$
- A_D = surface area of DRB on lightly doped side
- g = photogeneration rate function

A2 The n^+/p Junction

The n^+/p DR equations were originally listed in Reference [2] as

$$|U_p| = V_{DR} - V_T [1 - e^{-V_{DR}/V_T}] \quad (A1)$$

$$[N_A + V_T b \delta_o(j_e) + a_e \delta_o(j_e) V_T W / |U_p|] W^2 = 2\epsilon |U_p| / q \quad (A2)$$

$$\begin{aligned}
n_p^2 + [(1/2)N_A - N_D e^{-V_{DR}/V_T} - (3/2)V_{Tb} \delta_o(j_e)] n_p \\
- (1/2) [N_A - V_{Tb} \delta_o(j_e)] [V_{Tb} \delta_o(j_e) + N_D e^{-V_{DR}/V_T}] \\
- (1/4) [a_e \delta_o(j_e)]^2 (V_T / |U_p|) W^2 = 0 . \quad (A3)
\end{aligned}$$

Another equation is needed for j_h , which depends on the type of conditions considered (steady-state or transient, with or without photogeneration in the DR). The simplest case is steady-state with no photogeneration in the DR. For this case, it is often an adequate approximation to use $j_h \approx 0$ (a higher approximation is available [6] for use when this simple approximation is not adequate). To treat the steady-state case with photogeneration in the DR, we use the approximation that the hole current is negligible in the n^+ region adjacent to the metallurgical junction (MJ). Then the rate that holes move out of the DR through the p-side DRB is simply the photogeneration rate within the DR. Paying attention to the sign convention, the equation for j_h is

$$j_h = (q/A_D) \int_{DR} g d^3x = q W g_D \quad (A4)$$

where g is approximated by a constant value g_D within the DR. The complete list of n^+/p DR equations consists of (A1) through (A4).

When used to solve for W in terms of the other parameters, (A2) is a cubic equation. An exact analytic solution is available, but messy, and a simple approximation is more useful. To derive an approximation, we first simplify the notation in (A2) by defining

$$U \equiv |U_p|$$

$$S \equiv W/U^{1/2}$$

$$S_o \equiv (2\epsilon/q)^{1/2} [N_A + V_{Tb} j_e]^{-1/2}$$

$$K \equiv (q/2\epsilon) V_T a_e j_e .$$

We temporarily assume $j_e \geq 0$ so that $\delta_o(j_e) = 1$. The results applicable to $j_e < 0$ are trivial to derive and will be listed later. In this new notation, (A2) can be written as

$$[1/S_o^2 + K S/U^{1/2}] S^2 = 1 \quad (A5)$$

with S replacing W as the quantity to be solved. It is evident from (A5) that for fixed S_o and K , S has the asymptotic forms given by

$$S \rightarrow U^{1/6}/K^{1/3} \quad \text{as } U \rightarrow 0 \quad (A6a)$$

$$S \rightarrow S_o \quad \text{as } U \rightarrow \infty . \quad (A6b)$$

Write (A5) as

$$S = [1 - K S^3/U^{1/2}]^{1/2} S_o . \quad (A7)$$

The strategy is to replace the radical in (A7) with some approximation that makes the equation easy to solve for S . The asymptotic forms (A6) show that $KS^3/U^{1/2}$ varies between 0 and 1, so we look for an approximation for $[1-x^3]^{1/2}$ that is accurate for $0 \leq x \leq 1$. A particular approximation is

$$[1 - x^3]^{1/2} \approx [1 - x^{\sqrt{6}}]^{1/\sqrt{6}}$$

so that (A7) becomes

$$S/S_o \approx [1 - K^{\sqrt{6}/3} U^{-1/\sqrt{6}} S^{\sqrt{6}}]^{1/\sqrt{6}}$$

which can be solved for S with the result

$$S \approx [S_0^{-\sqrt{6}} + K^{\sqrt{6}/3} U^{-1/\sqrt{6}}]^{-1/\sqrt{6}} .$$

Going back to the original notation, the equation is written as

$$W \approx (2\epsilon/q)^{1/2} |U_p|^{1/2} [(N_A + V_T b j_e)^{\sqrt{6}/2} + (2\epsilon/q)^{1/\sqrt{6}} (V_T a_e j_e)^{\sqrt{6}/3} |U_p|^{-1/\sqrt{6}}]^{-1/\sqrt{6}} . \quad (A8)$$

Numerical calculations will show that (A8) is a very accurate approximation of (A2).

The original equation (A2) and the approximation (A8) both predict the same large $|U_p|$ behavior of W , which is

$$W \approx (2\epsilon/q)^{1/2} |U_p|^{1/2} [N_A + V_T b j_e]^{-1/2} \quad \text{for large } |U_p|. \quad (A9)$$

It is interesting that (A9) can also be derived by assuming velocity saturation in the DR (see any derivation of the Kirk effect). It should be expected that this assumption will lead to (A9) because velocity saturation is accompanied by large $|U_p|$. But the approximation breaks down for smaller $|U_p|$ and it is necessary to use the more accurate approximation (A8).

Equation (A3) can be greatly simplified, with a small accuracy penalty,¹ by replacing W with the large $|U_p|$ form given in (A9). The resulting equation can be solved for n_p in terms of V_{DR} (and j_e) but the equation is easier to write if V_{DR} is solved in terms of n_p instead of vice-versa. Solving for the exponential function gives

1. It is not clear that there is always an accuracy penalty, because (A3) is also only an approximation. Table 4.1 of Reference [2] compares n_p calculated from (A3) to the values calculated from a computer simulation for a few special cases. It turns out that, for these special cases, (A10) agrees better with the computer predictions than (A3).

$$\begin{aligned} \exp(-V_{DR}/V_T) &\approx N_D^{-1} [n_p - V_{Tb} j_e] \\ &- N_D^{-1} (V_T \epsilon/q) (a_e j_e)^2 [N_A \\ &+ V_{Tb} j_e]^{-1} [N_A - V_{Tb} j_e + 2n_p]^{-1} . \end{aligned} \quad (A10)$$

This equation restricts the possible values of n_p because the left side cannot be negative. The allowed values are bounded below by the asymptotic (large V_{DR}) limit, which is the largest value that makes the right side of (A10) zero.

The next approximation is to use $|U_p| \approx V_{DR}$ in (A8). In the original derivation, under steady-state conditions with no carrier generation in the DR, there was no distinction between the electron current at the DRB and at the MJ. But there is a distinction when carriers are generated in the DR, and it is best to use the electron current at the MJ in the DR equations. Neglecting the hole current at the MJ, the electron current at the MJ is the total current at the MJ, which is the total current at the DRB. The final modification to the DR equations is to replace j_e with $j_T = j_h + j_e$. The final results for the n^+/p DR, including those applicable to $j_T < 0$, are

$$\begin{aligned} W &= (2\epsilon/q)^{1/2} V_{DR}^{1/2} [(N_A + V_{Tb} j_T)^{v6/2} \\ &+ (2\epsilon/q)^{1/v6} (V_T a_e j_T)^{v6/3} V_{DR}^{-1/v6}]^{-1/v6} \quad \text{if } j_T \geq 0 \end{aligned}$$

$$W = [(2\epsilon/q) V_{DR}/N_A]^{1/2} \quad \text{if } j_T < 0$$

$$\begin{aligned} \exp(-V_{DR}/V_T) &= N_D^{-1} [n_p - V_{Tb} j_T] \\ &- N_D^{-1} (V_T \epsilon/q) (a_e j_T)^2 [N_A \\ &+ V_{Tb} j_T]^{-1} [N_A - V_{Tb} j_T + 2n_p]^{-1} \quad \text{if } j_T \geq 0 \end{aligned}$$

$$\exp(-V_{DR}/V_T) = N_D^{-1} n_p \quad \text{if } j_T < 0$$

$$j_h = (q/A_D) \int_{DR} g \, d^3x \quad .$$

A3 The p⁺/n Junction

The analogous equations for the p⁺/n DR are

$$W = (2\epsilon/q)^{1/2} V_{DR}^{1/2} [(N_D + V_T b j_T)^{6/2} + (2\epsilon/q)^{1/\sqrt{6}} (V_T a_h j_T)^{6/3} V_{DR}^{-1/\sqrt{6}}]^{-1/\sqrt{6}} \quad \text{if } j_T \geq 0$$

$$W = [(2\epsilon/q) V_{DR}/N_D]^{1/2} \quad \text{if } j_T < 0$$

$$\begin{aligned} \exp(-V_{DR}/V_T) = N_A^{-1} [p_n - V_T b j_T] \\ - N_A^{-1} (V_T \epsilon/q) (a_h j_T)^2 [N_D \\ + V_T b j_T]^{-1} [N_D - V_T b j_T + 2p_n]^{-1} \quad \text{if } j_T \geq 0 \end{aligned}$$

$$\exp(-V_{DR}/V_T) = N_A^{-1} p_n \quad \text{if } j_T < 0$$

$$j_e = (q/A_D) \int_{DR} g \, d^3x = q W g_D \quad .$$

APPENDIX B: THE SPECIAL FUNCTION H

B1 Introduction

Properties of the function H are discussed and a FORTRAN subprogram is provided for numerical evaluation. The subprogram can be appended to any FORTRAN driver code, allowing the code to call the function H as it would call any built-in function. Readers not interested in the analytical properties of H can read this introduction and then skip to Section B12 on page 134, where the subroutine can be found.

H is loosely defined by the equation

$$H(Z_1, Z_2) = E \quad \text{if and only if} \quad e^{1/E} = (E - Z_1)/(E - Z_2) . \quad (B1)$$

It is required that either no argument is negative or no argument is positive. It is also required that

$$1 + Z_1 - Z_2 \neq 0 \quad (B2)$$

although it is not obvious from a casual inspection of (B1) why (B2) is necessary. The function H has some subtle properties requiring a careful analysis, and even the definition has not yet been made sufficiently rigorous. Section B2 shows that (B1) sometimes makes sense, i.e., that there exists a unique value for $H(Z_1, Z_2)$ satisfying (B1) if Z_1 and Z_2 are suitably restricted. Sections B3 through B7 derive bounds, some of which are used in Section B8 to take some limits. These limits define H at some points that were excluded in Section B2. Nonnegative arguments are assumed until Section B10, which includes nonpositive arguments. Asymptotic forms are listed in Section B9, and a suggested algorithm for numerical evaluation is given in Section B11. A FORTRAN subprogram, using the suggested algorithm, is listed in Section B12.

B2 Definition of $H(Z_1, Z_2)$ when $Z_1 \geq 0$, $Z_2 > 0$, $Z_1 \neq Z_2$, and $1 + Z_1 - Z_2 \neq 0$

It is shown in this section that (B1) makes sense if

$$Z_1 \geq 0, \quad Z_2 > 0, \quad Z_1 \neq Z_2, \quad 1 + Z_1 - Z_2 \neq 0 \quad (\text{B3})$$

i.e., that there is a unique E satisfying

$$e^{1/E} = (E - Z_1)/(E - Z_2) \quad (\text{B4})$$

if (B3) is satisfied. Note that if we allowed the exponential function to have an infinite argument and if $Z_1=0$, we would call $E=0^-$ a solution to (B4), where the superscript means that E is on the negative side of zero (or $1/E = -\infty$). Also, if we allowed E to be infinite, we would call $E=\infty$ a solution. Such cases are not allowed and (B4) does not make sense if E is zero or infinite. Existence of a unique E satisfying (B4) means that there is a unique nonzero finite E satisfying (B4). The existence and uniqueness proof consists of two steps. The first step proves the existence and uniqueness of X_1 and X_2 satisfying the three conditions

$$(1 - X_1) e^{X_1} = (1 - X_2) e^{X_2} \quad (\text{B5a})$$

$$Z_1 X_2 - Z_2 X_1 = 1 + Z_1 - Z_2 \quad (\text{B5b})$$

$$X_1 \neq X_2 \quad (\text{B5c})$$

The second step uses the existence and uniqueness of the X 's to prove the existence and uniqueness of E satisfying (B4).

Before carrying out the first step, it is necessary to define and establish some properties of a particular function g (not to be confused with the generation rate function). We start with the function f defined by

$$f(X) \equiv (1 - X) e^X . \quad (B6)$$

Differentiating gives $f'(X) = -Xe^X$, so f is strictly increasing when X is negative and strictly decreasing when X is positive. Therefore, f is invertible on each branch, i.e., there is an f_1^{-1} and f_2^{-1} satisfying

$$f_1^{-1}(f(X)) = X \quad \text{if } X < 0 \quad (B7a)$$

$$f_2^{-1}(f(X)) = X \quad \text{if } 0 < X < 1 . \quad (B7b)$$

Some mapping properties are

$$f : (-\infty, 0) \leftrightarrow (0, 1) , \quad f_1^{-1} : (0, 1) \leftrightarrow (-\infty, 0) \quad (B8a)$$

$$f : (0, 1) \leftrightarrow (0, 1) , \quad f_2^{-1} : (0, 1) \leftrightarrow (0, 1) \quad (B8b)$$

where \leftrightarrow means that the mapping is one-to-one and onto. The function f maps the two intervals $(-\infty, 0)$ and $(0, 1)$ onto the same target set $(0, 1)$, which is the domain of both inverses. Note that both inverses are right inverses, i.e.,

$$f(f_1^{-1}(Y)) = Y \quad \text{and} \quad f(f_2^{-1}(Y)) = Y \quad \text{if } 0 < Y < 1 . \quad (B9)$$

The function g is defined by

$$g(0) \equiv 0 \quad (B10a)$$

$$g(X) \equiv f_2^{-1}(f(X)) \quad \text{if } X < 0 \quad (B10b)$$

$$g(X) \equiv f_1^{-1}(f(X)) \quad \text{if } 0 < X < 1 \quad (B10c)$$

and (B9) together with $f(0)=1$ gives

$$f(g(X)) = f(X) \quad \text{if } X < 1 . \quad (\text{B11})$$

Using (B8) and (B10), we find that g has the mapping properties

$$g : (-\infty, 0) \leftrightarrow (0, 1) \quad (\text{B12a})$$

$$g : (0, 1) \leftrightarrow (-\infty, 0) . \quad (\text{B12b})$$

The function g is easiest to visualize from (B11) and (B12). Given that $X < 1$ and $X \neq 0$, we can think of $g(X)$ as "the other X producing the same $f(X)$." In other words, $f(g(X))=f(X)$ but $g(X) \neq X$. In fact,

if $X < 1$ and $X \neq 0$, then $g(X) \neq 0$ and $g(X)$ is negative
(positive) if and only if X is positive (negative). (B13)

By combining (B10a) with (B12), we get

$$g : (-\infty, 1) \leftrightarrow (-\infty, 1) . \quad (\text{B14})$$

Note that f is strictly increasing on $(-\infty, 0)$ and decreasing on $(0, 1)$, so f_1^{-1} is increasing and f_2^{-1} is decreasing. From (B10), we conclude that g is decreasing on $(-\infty, 0)$ and on $(0, 1)$. But g is continuous at $X=0$ and we conclude

$$g \text{ is strictly decreasing on } (-\infty, 1) . \quad (\text{B15})$$

Having established some properties of g , we can now show that there exists a unique X_1 and X_2 satisfying (B5). Note that (B5a) and (B5c) together imply that X_1 and X_2 are both less than 1.

This can be shown by contradiction. Assume $X_2 \geq 1$. Then the right side of (B5a) is negative or zero. No negative X_1 can make the left side negative or zero, so X_1 must be positive or zero. But invertibility of f on $[0, \infty)$ contradicts (B5c). Therefore, (B5) implies that $X_2 < 1$. Similarly, $X_1 < 1$. Using the definition of g , we can write (B5a) as $X_2 = g(X_1)$. Using this to eliminate X_2 in (B5b), we find that (B5) implies that

$$Z_1 g(X_1) - Z_2 X_1 = 1 + Z_1 - Z_2 \quad (\text{B16a})$$

$$X_2 = g(X_1) \quad (\text{B16b})$$

Conversely, (B16) implies (B5). The equation (B16b) implies (B5a) and $X_1, X_2 < 1$, and the two equations in (B16) imply (B5b). Furthermore, $X_1 \neq 0$ because the right side of (B16a) is not zero (by assumption (B3)). Using (B13), we conclude (B5c). Therefore, (B5) and (B16) are equivalent. To show that there is a unique X_1 and X_2 satisfying (B5), it suffices to show that there is a unique X_1 and X_2 satisfying (B16). Note that g is strictly decreasing and (by assumption (B3)), $Z_1 \geq 0$ and $Z_2 > 0$. Therefore the left side of (B16a), regarded as a function of X_1 , is strictly decreasing. Therefore an X_1 satisfying (B16a) is unique if it exists. The left side maps $(-\infty, 1)$ onto $(-\infty, \infty)$ if $Z_1 > 0$. If $Z_1 = 0$, the left side maps $(-\infty, 1)$ onto $(-Z_2, \infty)$. In either case, the target set includes the point $1 + Z_1 - Z_2$. Therefore there is a unique X_1 satisfying (B16a). There is a unique X_2 satisfying (B16b) and this completes the proof of existence and uniqueness for the X 's satisfying (B5).

We next prove the existence and uniqueness of E satisfying (B4). To prove existence, start with the X_1 and X_2 satisfying (B5) and let

$$E = 1/(X_2 - X_1) \quad (\text{B17})$$

X_1 and X_2 exist (are finite) so $E \neq 0$. Furthermore, $X_1 \neq X_2$ so E given by (B17) exists (is finite). By assumption (B3), $Z_1 \neq Z_2$, so (B5b) and (B17) can be used to solve for the X 's in terms of E with the result

$$X_1 = [Z_1 - (1 + Z_1 - Z_2) E] / [(Z_2 - Z_1) E] \quad (B18a)$$

$$X_2 = [Z_2 - (1 + Z_1 - Z_2) E] / [(Z_2 - Z_1) E] . \quad (B18b)$$

Substituting (B18) into (B5a) shows that E satisfies (B4), which establishes existence of a solution to (B4). Uniqueness is proven by reversing these steps. Let E satisfy (B4) and define X_1 and X_2 by (B18). Using (B18) together with (B4) shows that the X's satisfy (B5), implying that the X's are unique. But E is also related to the X's by (B17), implying that E is unique.

This completes the proof of existence and uniqueness of E. We define $H(Z_1, Z_2)$ to be this E, so it is now defined for all Z_1 and Z_2 satisfying (B3).

B3 Some Inequalities

The function H has been defined when the Z's satisfy (B3). Some other cases such as $Z_1=Z_2$ or $Z_2=0$ violate (B3), and limits will be used to define $H(Z_1, Z_2)$ for those cases. Some bounds for $H(Z_1, Z_2)$ will help to evaluate these limits. The first step, and the objective of this section, is to derive bounds for the X's satisfying (B5). These bounds will then be used in the next three sections to derive bounds for $H(Z_1, Z_2)$. The bounds for the X's derived here will also be used by the numerical algorithm in Section B11. It is assumed throughout this section that the Z's satisfy (B3) so that E satisfying (B4) is related by (B18) to the X's satisfying (B5).

It was concluded in the previous section that the X's are both less than 1 and that X_1 is not zero. By combining (B13) with (B16b), we conclude that X_2 is not zero and the two X's have opposite signs. Therefore, one of the X's is negative and the other is positive and in the interval (0,1). Equation (B5b) can be used to identify which of the two X's is negative, and the first pair of inequalities is

$$X_1 < 0 \quad \text{and} \quad 0 < X_2 < 1 \quad \text{if} \quad 1 + Z_1 - Z_2 > 0 \quad (B19a)$$

$$0 < X_1 < 1 \quad \text{and} \quad X_2 < 0 \quad \text{if} \quad 1 + Z_1 - Z_2 < 0 \quad . \text{(B19b)}$$

Other bounds can be obtained by substituting (B19) back into (B5b). For example, if $1+Z_1-Z_2>0$, then $X_1<0$ which implies a bound on X_2 via (B5b), i.e.,

$$Z_1 X_2 < 1 + Z_1 - Z_2 \quad \text{if} \quad 1 + Z_1 - Z_2 > 0 \quad . \quad \text{(B20a)}$$

Similarly,

$$Z_1 X_2 > 1 + Z_1 - Z_2 \quad \text{if} \quad 1 + Z_1 - Z_2 < 0 \quad . \quad \text{(B20b)}$$

To obtain (sometimes) tighter bounds, we need a tool derived by differentiating $e^X - (1+X)$ to conclude that the expression is minimum at $X=0$. The expression is larger at any $X \neq 0$ than at $X=0$, or

$$e^X - (1 + X) > 0 \quad \text{if} \quad X \neq 0 \quad . \quad \text{(B21)}$$

Now differentiate the expression $e^X - (1+X+X^2/2)$ and use (B21) to conclude that the expression is strictly increasing. The expression is larger at any $X>0$ than at $X=0$, and smaller at any $X<0$ than at $X=0$. This gives

$$e^X > 1 + X + X^2/2 \quad \text{if} \quad X > 0 \quad \text{(B22a)}$$

$$e^X < 1 + X + X^2/2 \quad \text{if} \quad X < 0 \quad . \quad \text{(B22b)}$$

Now write (B5a) as

$$(1 - X_1) e^{-X_2} = (1 - X_2) e^{-X_1} \quad . \quad \text{(B23)}$$

First assume that $X_1 < 0$, implying that $X_2 > 0$. Applying (B22a) to the right side of (B23), and (B22b) to the left side and rearranging terms gives

$$(1 - X_1) (1 - X_2) < 1 . \quad (B24)$$

When deriving (B24), it was assumed that X_1 is negative and X_2 positive. Interchanging indices for the other case produces the same result, so (B24) applies to all cases. Note that (B24) can be manipulated into

$$X_1 / (1 - X_1) > - X_2 \quad (\text{equivalent to (B24)}) . \quad (B25)$$

To derive another inequality, note that (B5a) defines X_2 as a function of X_1 . Differentiating (B5a) and then using (B5a) again to eliminate the exponential function gives

$$dX_2/dX_1 = [(1 - X_2) X_1] / [(1 - X_1) X_2] .$$

First assume that $X_1 < 0$, implying that $0 < X_2 < 1$. Combining (B25) with the above equation gives

$$dX_2/dX_1 > X_2 - 1 > - 1 .$$

The direction of the inequality is preserved upon integration if the upper integration limit is larger than the lower. Integrating from an arbitrary $X_1 < 0$ to $X_1 = 0$ while using $X_2 \rightarrow 0$ as $X_1 \rightarrow 0$ gives

$$X_1 + X_2 < 0 . \quad (B26)$$

When deriving (B26), it was assumed that $X_1 < 0$. Interchanging indices produces the same result, so (B26) applies to all cases. Note that (B26) states that the negative X has the larger absolute value.

The inequalities (B24) and (B26) were derived from (B5a) alone. Including (B5b) allows (B26) to be written as

$$X_2 < (1 + Z_1 - Z_2)/(Z_1 + Z_2)$$

and (B24) can be written as

$$Z_1 (1 - X_2)^2 + (1 - X_2) < Z_2 .$$

Temporarily assuming that $Z_1 \neq 0$, this inequality can be manipulated into

$$[(1 - X_2) + 1/(2Z_1)]^2 < Z_2/Z_1 + 1/(2Z_1)^2 .$$

The expression in brackets is positive (because $X_2 < 1$), so taking the square root and rearranging terms gives

$$X_2 > 2(1 + Z_1 - Z_2) [1 + 2Z_1 + (1 + 4Z_1 Z_2)^{1/2}]^{-1}$$

which is also valid when $Z_1 = 0$.

The important results when (B3) applies, excluding inequalities that are always implied by others, are listed below for X_2 (corresponding bounds for X_1 are implied by (B5b)):

$$X_2 > 0 \quad \text{if} \quad 1 + Z_1 - Z_2 > 0 \quad (\text{B27a})$$

$$X_2 < 1 \quad (\text{B27b})$$

$$Z_1 X_2 > 1 + Z_1 - Z_2 \quad \text{if} \quad 1 + Z_1 - Z_2 < 0 \quad (\text{B27c})$$

$$X_2 < (1 + Z_1 - Z_2)/(Z_1 + Z_2) \quad (\text{B27d})$$

$$X_2 > 2(1 + Z_1 - Z_2) [1 + 2Z_1 + (1 + 4Z_1 Z_2)^{1/2}]^{-1} . \quad (\text{B27e})$$

B4 Bounds for Case 1: $0 < Z_2 < Z_1$

We assume that (B3) is satisfied and derive some bounds for $H(Z_1, Z_2)$, which is E satisfying (B4). It is convenient to consider several cases separately. We start with Case 1, defined by the condition

$$0 < Z_2 < Z_1 \quad (\text{defines Case 1}) . \quad (\text{B28})$$

E can be expressed in terms of X_2 via (B18b), with the result

$$E = [(Z_2 - Z_1) X_2 + 1 + Z_1 - Z_2]^{-1} Z_2 \quad (\text{B29})$$

so that bounds for X_2 imply corresponding bounds for E. Using the applicable inequalities in (B27), and paying attention to the fact that $Z_1 - Z_2$ and $1 + Z_1 - Z_2$ are positive (for Case 1) when rearranging terms, gives

$$E < Z_2$$

$$E < (1/2) (Z_1 + Z_2) / (1 + Z_1 - Z_2) \quad (\text{B30})$$

$$E > Z_2 / (1 + Z_1 - Z_2) > 0 . \quad (\text{B31})$$

Another bound can be obtained by writing (B4) as

$$E = Z_2 - (Z_1 - Z_2) / (e^{1/E} - 1) .$$

The expression on the right, regarded as a function of E, is strictly decreasing (if E>0), so it maps upper bounds into lower bounds and positive lower bounds into upper bounds. Using the upper bound Z_2 , we obtain the lower bound

$$E > Z_2 - (Z_1 - Z_2)/(e^{1/Z_2} - 1) .$$

Still more bounds can be obtained by writing (B4) as

$$E = \{\ln[(E - Z_1)/(E - Z_2)]\}^{-1} . \quad (B32)$$

Because $Z_1 > Z_2$, the right side of (B32), regarded as a function of E, is strictly decreasing on the interval $(0, Z_2)$. Therefore the right side of (B32) maps lower bounds for E into upper bounds and upper bounds into lower bounds, if E and the original bounds are in the required interval $(0, Z_2)$. But E and the lower bound in (B31) are in the required interval, and we obtain the new upper bound

$$E < \{\ln[(Z_1 + 1)/Z_2]\}^{-1} .$$

The upper bound in (B30) will be in the required interval if $Z_2 > 1/2$, and we obtain the new lower bound

$$E > \{\ln[(Z_1 + 1/2)/(Z_2 - 1/2)]\}^{-1} \quad \text{if } Z_2 > 1/2 .$$

The bounds for $H(Z_1, Z_2)$ (=E) are summarized below.

If $0 < Z_2 < Z_1$ (Case 1), then:

$$H(Z_1, Z_2) < Z_2 \quad (B33a)$$

$$H(Z_1, Z_2) < (1/2) (Z_1 + Z_2)/(1 + Z_1 - Z_2) \quad (B33b)$$

$$H(Z_1, Z_2) < \{\ln[(Z_1 + 1)/Z_2]\}^{-1} \quad (\text{B33c})$$

$$H(Z_1, Z_2) > Z_2/(1 + Z_1 - Z_2) > 0 \quad (\text{B33d})$$

$$H(Z_1, Z_2) > Z_2 - (Z_1 - Z_2)/(e^{1/Z_2} - 1) \quad (\text{B33e})$$

$$H(Z_1, Z_2) > \{\ln[(Z_1 + 1/2)/(Z_2 - 1/2)]\}^{-1} \quad \text{if } Z_2 > 1/2. \quad (\text{B33f})$$

B5 Bounds for Case 2: $0 \leq Z_1 < Z_2 < Z_1 + 1$

Case 2 is defined by the condition

$$0 \leq Z_1 < Z_2 < Z_1 + 1 \quad (\text{defines Case 2}) \quad (\text{B34})$$

Some of the applicable inequalities in (B27) combined with (B29) give

$$0 < Z_2 < E$$

$$(1/2) (Z_1 + Z_2)/(1 + Z_1 - Z_2) < E < Z_2/(1 + Z_1 - Z_2). \quad (\text{B35})$$

Another bound is obtained by using (B18) to write (B24) as

$$(E - Z_1) (E - Z_2) < (Z_2 - Z_1)^2 E^2$$

or

$$(1+Z_1-Z_2) (1-Z_1+Z_2) E^2 < (Z_1+Z_2) E - Z_1 Z_2 \leq (Z_1+Z_2) E.$$

Dividing by the positive quantity $(1+Z_1-Z_2)(1-Z_1+Z_2)E$ gives

$$E < (Z_1 + Z_2) / [(1 + Z_1 - Z_2) (1 - Z_1 + Z_2)] .$$

Because $Z_2 > Z_1$, the right side of (B32), regarded as a function of E , is strictly increasing on the interval $(Z_2, +\infty)$, which contains E and both bounds in (B35) if $Z_2 > 1/2$. Therefore, the right side of (B32) maps the upper bound into an upper bound and (if $Z_2 > 1/2$) the lower bound into a lower bound. The new bounds are the same as obtained for Case 1. The bounds for $H(Z_1, Z_2)$ ($=E$) are summarized below.

If $0 \leq Z_1 < Z_2 < Z_1 + 1$ (Case 2), then:

$$H(Z_1, Z_2) < Z_2 / (1 + Z_1 - Z_2) \quad (\text{B36a})$$

$$H(Z_1, Z_2) < (Z_1 + Z_2) / [(1 + Z_1 - Z_2) (1 - Z_1 + Z_2)] \quad (\text{B36b})$$

$$H(Z_1, Z_2) < \{\ln[(Z_1 + 1)/Z_2]\}^{-1} \quad (\text{B36c})$$

$$H(Z_1, Z_2) > Z_2 > 0 \quad (\text{B36d})$$

$$H(Z_1, Z_2) > (1/2) (Z_1 + Z_2) / (1 + Z_1 - Z_2) \quad (\text{B36e})$$

$$H(Z_1, Z_2) > \{\ln[(Z_1 + 1/2)/(Z_2 - 1/2)]\}^{-1} \quad \text{if } Z_2 > 1/2 . \quad (\text{B36f})$$

B6 Bounds for Case 3: $1 \leq Z_1 + 1 < Z_2$

Case 3 is defined by the condition

$$1 \leq Z_1 + 1 < Z_2 \quad (\text{defines Case 3}) \quad (\text{B37})$$

Some of the applicable inequalities in (B27) combined with (B29) give

$$(1/2) (Z_1 + Z_2)/(1 + Z_1 - Z_2) < E < Z_1/(1 + Z_1 - Z_2) \leq 0 \text{ .(B38)}$$

Steps similar to those that produced (B36b) give

$$E < (Z_1 + Z_2)/[(1 + Z_1 - Z_2) (1 - Z_1 + Z_2)] \text{ .}$$

Because $Z_2 > Z_1$, the right side of (B32), regarded as a function of E , is strictly increasing on $(-\infty, 0)$, which contains E and both bounds in (B38) if $Z_1 > 0$. Therefore the right side of (B32) maps the lower bound into a lower bound and (if $Z_1 > 0$) the upper bound into an upper bound. The bounds for $H(Z_1, Z_2)$ ($=E$) are summarized below.

If $1 \leq Z_1 + 1 < Z_2$ (Case 3), then:

$$H(Z_1, Z_2) < Z_1/(1 + Z_1 - Z_2) \leq 0 \quad (\text{B39a})$$

$$H(Z_1, Z_2) < (Z_1 + Z_2)/[(1 + Z_1 - Z_2) (1 - Z_1 + Z_2)] < 0 \quad (\text{B39b})$$

$$H(Z_1, Z_2) < \{\ln[Z_1/(Z_2 - 1)]\}^{-1} \quad \text{if } Z_1 > 0 \quad (\text{B39c})$$

$$H(Z_1, Z_2) > (1/2) (Z_1 + Z_2)/(1 + Z_1 - Z_2) \quad (\text{B39d})$$

$$H(Z_1, Z_2) > \{\ln[(Z_1 + 1/2)/(Z_2 - 1/2)]\}^{-1} \text{ .} \quad (\text{B39e})$$

B7 Some Additional Bounds for X2

The bounds for E (=H(Z₁,Z₂)) derived in the last three sections are adequate for the intended purpose of determining a few selected limits and asymptotic forms (next two sections). But unless a limit or asymptotic form is found to apply, numerical evaluation of H(Z₁,Z₂) will work with the X's and the number of required calculations is reduced by tightening the bounds for X₂. It is therefore desirable to use all available information to bracket X₂ as tightly as possible. Some of the bounds for E are equivalent to (via (B18b)) or weaker than the X₂ bounds in (B27). Some other E bounds, such as (B33c), were obtained from an additional step and can be used to derive new X₂ bounds. Using (B18b), (B33c), (B33e), (B33f), (B36c), (B36f), (B39c), and (B39e) provides the following additional bounds:

$$X_2 < \frac{1+Z_1-Z_2}{Z_1-Z_2} + \frac{Z_2}{Z_2-Z_1} \ln \left[\frac{Z_1+1}{Z_2} \right] \quad \text{if } 0 < Z_2 < Z_1$$

$$X_2 < \frac{1+Z_1-Z_2}{Z_1-Z_2} + \frac{Z_2}{Z_2-Z_1} \ln \left[\frac{Z_1+1/2}{Z_2-1/2} \right] \quad \text{if } Z_1 \geq 0, Z_2 > 1/2, Z_2 > Z_1 \\ \text{and } 1+Z_1-Z_2 \neq 0$$

$$X_2 > \frac{1+Z_1-Z_2}{Z_1-Z_2} + \frac{Z_2}{Z_2-Z_1} \ln \left[\frac{Z_1+1/2}{Z_2-1/2} \right] \quad \text{if } 1/2 < Z_2 < Z_1$$

$$X_2 > \frac{1+Z_1-Z_2}{Z_1-Z_2} + \frac{Z_2}{Z_2-Z_1} \ln \left[\frac{Z_1+1}{Z_2} \right] \quad \text{if } 0 \leq Z_1 < Z_2 < Z_1+1$$

$$X_2 > \frac{1+Z_1-Z_2}{Z_1-Z_2} + \frac{Z_2}{Z_2-Z_1} \ln \left[\frac{Z_1}{Z_2-1} \right] \quad \text{if } 1 < Z_1+1 < Z_2$$

$$X_2 > \frac{1+Z_1-Z_2}{Z_1-Z_2} + \frac{Z_2}{Z_2-Z_1} \frac{1 - e^{-1/Z_2}}{Z_2 - Z_1 e^{-1/Z_2}} \quad \text{if } 0 < Z_2 < Z_1 \text{ and } Z_2 < [\ln(Z_1/Z_2)]^{-1}$$

B8 Definition of H(Z,Z) and H(Z,0) when Z ≥ 0

The quantities H(Z,Z) and H(Z,0) are not yet defined because the arguments violate (B3). These quantities will be defined, with Z ≥ 0, by taking limits. A limit of a function of several variables can be subtle because a given point can be approached along a variety of paths, and the limit is well defined only if all possible paths produce the same limit. Fortunately, the limits needed here are well defined.

First consider the limit as (Z₁, Z₂) approaches (Z, Z) for some Z ≥ 0. If (Z₁, Z₂) is sufficiently close to (Z, Z), Case 3 is excluded and the bounds (B33a), (B33d), (B36a), and (B36d) imply that, no matter what path is followed, we have H(Z₁, Z₂) → Z. We define H(Z, Z) to be this limit, i.e.,

$$H(Z, Z) \equiv Z \quad \text{if } Z \geq 0. \quad (\text{B40a})$$

Now consider the limit as (Z₁, Z₂) approaches (Z, 0) for some Z ≥ 0. We may assume that Z > 0, because (B40) applies if Z = 0. But if Z > 0 and (Z₁, Z₂) is sufficiently close to (Z, 0), only Case 1 can apply. The bounds (B33a) and (B33d) imply that, no matter what path is followed, we have H(Z₁, Z₂) → 0. We define H(Z, 0) to be this limit, i.e.,

$$H(Z, 0) \equiv 0 \quad \text{if } Z \geq 0. \quad (\text{B40b})$$

The condition Z = 0 was allowed in (B40b) because (B40a) and (B40b) are equivalent when Z = 0. The quantity H(Z₁, Z₂) is now defined for all nonnegative Z₁ and Z₂ satisfying 1 + Z₁ - Z₂ ≠ 0.

B9 Asymptotic Forms

Asymptotic forms are approximations for $H(Z_1, Z_2)$ that become exact (in the sense that the relative or fractional error goes to zero) in the limit as various combinations of the arguments become small or large. Such approximations make the behavior of H easier to visualize. They also have computational advantages (when applicable) because they are simple. With (B3) assumed, asymptotic forms are derived below for: small $|Z_1 - Z_2|$, small Z_2 , small $|1 + Z_1 - Z_2|$, and large Z_1 and Z_2 . Applicability tests are given in terms of a positive quantity δ_{tol} , which is a user specified relative error that will be tolerated in the estimate of $H(Z_1, Z_2)$. For example, if an error less than one percent is good enough, whether too large or too small, then $\delta_{tol} = 0.01$.

The first asymptotic form applies when $\epsilon_1 = |Z_1 - Z_2|$ is small. A sufficiently small ϵ_1 excludes Case 3. To insure that Case 3 is excluded, it is required that $\epsilon_1 < 1$. For Case 1 conditions, $\epsilon_1 = Z_1 - Z_2$ and the bounds (B33a) and (B33d) can be written as

$$Z_2 / (1 + \epsilon_1) < H(Z_1, Z_2) < Z_2$$

Similarly, Case 2 gives

$$Z_2 < H(Z_1, Z_2) < Z_2 / (1 - \epsilon_1) .$$

In either case, the approximation $H(Z_1, Z_2) \approx Z_2$ has a relative error less than ϵ_1 , if $\epsilon_1 < 1$. The approximation has a relative error less than δ_{tol} if $\epsilon_1 < 1$ and $\epsilon_1 < \delta_{tol}$, i.e.,

$H(Z_1, Z_2) \approx Z_2$ has relative error less than δ_{tol} if (B3) (B41a) applies and $|Z_1 - Z_2| < 1$ and $|Z_1 - Z_2| < \delta_{tol}$.

The second asymptotic form applies when Z_2 is small. A sufficiently small Z_2 excludes Case 3 conditions. A small Z_2 under

Case 2 conditions implies a small $|Z_1 - Z_2|$ and (B41a) applies. It is therefore adequate to consider only Case 1 conditions by requiring that $0 < Z_2 < Z_1$. "Small Z_2 " under Case 1 conditions is interpreted to mean that ϵ_2 is small, where

$$\epsilon_2 \equiv [(Z_1 - Z_2)/Z_2] [e^{1/Z_2} - 1]^{-1} .$$

The Case 1 bounds (B33a) and (B33e) can be written as

$$(1 - \epsilon_2) Z_2 < H(Z_1, Z_2) < Z_2 .$$

In addition to $0 < Z_2 < Z_1$, we also require that the Z's satisfy the condition $\epsilon_2 < 1$. Then the above bounds imply that the approximation $H(Z_1, Z_2) \approx Z_2$ has a relative error less than $\epsilon_2/(1 - \epsilon_2)$. The approximation has a relative error less than δ_{tol} if the requirements $0 < Z_2 < Z_1$, $\epsilon_2 < 1$, and $\epsilon_2/(1 - \epsilon_2) < \delta_{tol}$ are all satisfied. Using the definition of ϵ_2 to express the ϵ_2 requirements in terms of the Z's and noting that one of the inequalities is implied by the other, we obtain

$$H(Z_1, Z_2) \approx Z_2 \quad \text{has relative error less than } \delta_{tol} \quad \text{if} \quad (B41b)$$

$$(B3) \text{ applies and } Z_2 < Z_1 \text{ and}$$

$$Z_2 < \{\ln[Z_1/Z_2 + (Z_1 - Z_2)/(Z_2 \delta_{tol})]\}^{-1} .$$

The third asymptotic form applies when the Z's come close to the forbidden condition $1 + Z_1 - Z_2 = 0$, i.e., $\epsilon_3 \equiv |1 + Z_1 - Z_2|$ is small. A sufficiently small ϵ_3 excludes Case 1 conditions. We insure that Case 1 is excluded by requiring that $\epsilon_3 < 1$. For Case 2, $\epsilon_3 = 1 + Z_1 - Z_2$ and (B36b) and (B36e) can be written as

$$[(1/2) (Z_1 + Z_2)/(1 + Z_1 - Z_2)] < H(Z_1, Z_2)$$

$$< [(1/2) (Z_1 + Z_2)/(1 + Z_1 - Z_2)] / (1 - \epsilon_3/2) .$$

Similarly, Case 3 gives

$$\begin{aligned} & [- (1/2) (Z_1 + Z_2)/(1 + Z_1 - Z_2)] / (1 + \epsilon_3/2) < - H(Z_1, Z_2) \\ & < [- (1/2) (Z_1 + Z_2)/(1 + Z_1 - Z_2)] . \end{aligned}$$

In either case, the approximation

$$H(Z_1, Z_2) \approx (1/2) (Z_1 + Z_2)/(1 + Z_1 - Z_2)$$

has a relative error less than $\epsilon_3/2$, if $\epsilon_3 < 1$. The approximation has a relative error less than δ_{tol} if $\epsilon_3 < 1$ and $\epsilon_3/2 < \delta_{tol}$, i.e.,

$$H(Z_1, Z_2) \approx (1/2) (Z_1 + Z_2)/(1 + Z_1 - Z_2)$$

has relative error less than δ_{tol} if (B41c)
 (B3) applies and
 $|1+Z_1-Z_2| < \min\{ 1 , 2\delta_{tol} \}$.

The last asymptotic form applies when the Z's are both large in the sense that the upper and lower logarithmic bounds ((B33c) and (B33f) for Cases 1 and 2, and (B39c) and (B39e) for Case 3) come together. An equivalent statement is that ϵ_4 is small, where

$$\epsilon_4 \equiv 1 - \ln[(Z_1 + 1)/Z_2] / \ln[(Z_1 + 1/2)/(Z_2 - 1/2)]$$

$$\text{if } Z_2 > 1/2 \text{ and } 1 + Z_1 - Z_2 > 0$$

$$\epsilon_4 \equiv \ln[Z_1/(Z_2 - 1)] / \ln[(Z_1 + 1/2)/(Z_2 - 1/2)] - 1$$

$$\text{if } Z_1 > 0 \text{ and } 1 + Z_1 - Z_2 < 0 .$$

The logarithmic bounds can be written as

$$\{1/\ln[(Z_1 + 1/2)/(Z_2 - 1/2)]\} < H(Z_1, Z_2)$$

$$< \{1/\ln[(Z_1 + 1/2)/(Z_2 - 1/2)]\}/(1 - \epsilon_4)$$

$$\text{if } Z_1 \geq 0, \quad Z_2 > 1/2, \quad Z_2 \neq Z_1, \quad \text{and } 1 + Z_1 - Z_2 > 0$$

$$\{-1/\ln[(Z_1 + 1/2)/(Z_2 - 1/2)]\}/(1 + \epsilon_4) < -H(Z_1, Z_2)$$

$$< \{-1/\ln[(Z_1 + 1/2)/(Z_2 - 1/2)]\}$$

$$\text{if } Z_1 > 0 \quad \text{and} \quad 1 + Z_1 - Z_2 < 0.$$

In either case, the approximation

$$H(Z_1, Z_2) \approx \{\ln[(Z_1 + 1/2)/(Z_2 - 1/2)]\}^{-1}$$

has a relative error less than ϵ_4 . The relative error will be less than δ_{tol} if $\epsilon_4 < \delta_{\text{tol}}$. The fact that $\ln(a+b) < \ln(a) + b/a$, when $a > 0$ and $b > 0$, can be used to show that if $Z_1 > 0$, $Z_2 > 1/2$, and $1 + Z_1 - Z_2 \neq 0$, then

$$\epsilon_4 < (1/2Z_1) [(1+Z_1-Z_2)/(Z_2-1/2)] \{\ln[(Z_1+1/2)/(Z_2-1/2)]\}^{-1}$$

so the relative error is less than δ_{tol} if the Z's satisfy the stated conditions and the right side of the above inequality is less than δ_{tol} . This gives

$$H(Z_1, Z_2) \approx \{\ln[(Z_1 + 1/2)/(Z_2 - 1/2)]\}^{-1}$$

has relative error less than δ_{tol} if (B41d)
(B3) applies and $Z_1 > 0$, $Z_2 > 1/2$, and

$$1/Z_1 < 2\delta_{\text{tol}} [(Z_2-1/2)/(1+Z_1-Z_2)] \ln[(Z_1+1/2)/(Z_2-1/2)].$$

B10 Definition of H(Z1,Z2) when Z1<0, Z2<0, and 1+Z1-Z2≠0

So far, $H(Z_1, Z_2)$ has been defined for any nonnegative Z's satisfying $1+Z_1-Z_2 \neq 0$. The definition is either (B1) or limits derived from it. A similar definition can be given when the arguments are nonpositive. For this case, we can manipulate (B1) into

$$e^{1/(-E)} = [(-E) - (-Z_2)]/[(-E) - (-Z_1)]$$

which has the same solutions and limits as the original (B1), except for a change in symbols; $-E$ replaces E , $-Z_2$ replaces Z_1 , and $-Z_1$ replaces Z_2 . The problem case is $1+(-Z_2)-(-Z_1)=0$, so even for nonpositive arguments the problem case is still $1+Z_1-Z_2=0$. We therefore define

$$H(Z_1, Z_2) \equiv -H(-Z_2, -Z_1) \quad \text{if } Z_1 < 0, \quad Z_2 < 0, \quad \text{and } 1+Z_1-Z_2 \neq 0 \quad (\text{B42})$$

so $H(Z_1, Z_2)$ is now defined for all Z's satisfying $Z_1 Z_2 \geq 0$ and $1+Z_1-Z_2 \neq 0$.

B11 A Numerical Algorithm

This section suggests one possible algorithm for numerical evaluation of $H(Z_1, Z_2)$. Because of (B42), it is sufficiently general to confine our attention to nonnegative arguments. It is assumed below that the Z's are nonnegative and satisfy $1+Z_1-Z_2 \neq 0$.

The first step is to determine the applicability of the special cases (B40) and asymptotic forms (B41) in the order listed. Use the first case that was found to apply. If none apply, then the Z's satisfy (B3), implying that $H(Z_1, Z_2)$ is the E satisfying (B4) and can be solved via (B29) from the X's satisfying (B5). The equations governing the X's can be written as

$$X_1 = [Z_1 X_2 - (1 + Z_1 - Z_2)]/Z_2 \quad (\text{B43a})$$

$$W \equiv X_1 - X_2 + \ln[(1 - X_1)/(1 - X_2)] \quad (\text{B43b})$$

$$W = 0 .$$

Let X_A and X_B be, respectively, the largest lower bound and smallest upper bound for X_2 that can be found in (B27) and in Section B7. The basic idea is to guess at X_2 , then calculate X_1 from (B43) and W from (B44). Whether the guess is too large or too small depends on the sign of W . The guess is too large if W has the same sign as W_B , where W_B is the value obtained when X_2 is replaced with X_B in (B43). The guess is too small if W and W_B have opposite signs, and the guess is correct if $W=0$. The bisection method is used to construct a sequence of lower X_2 bounds $X_A^{(1)}, X_A^{(2)}, \dots$, and upper bounds $X_B^{(1)}, X_B^{(2)}, \dots$. The first bounds are $X_A^{(1)}=X_A, X_B^{(1)}=X_B$. For $i>1$, the i^{th} bounds are constructed from the previous bounds by letting $X_M^{(i-1)}$ be the midpoint

$$X_M^{(i-1)} = (1/2)[X_A^{(i-1)} + X_B^{(i-1)}] .$$

Determine the sign of W obtained when X_2 is replaced with this midpoint. If W and W_B have the same sign, the correct X_2 is smaller than the midpoint and the new lower bound is the same as the old while the new upper bound is the old midpoint. Similarly, if W and W_B have opposite signs, the new lower bound is the old midpoint and the upper bound is not changed. As the upper and lower X_2 bounds come together, the corresponding E bounds (from (B29)) also come together. The bisection is terminated when the E bounds are sufficiently close together.

B12 The Function Subprogram

The subprogram listed at the end of this discussion can be appended to any FORTRAN source code, allowing the code to call the function H as it would call any built-in function. The subprogram uses the numerical algorithm discussed in the previous section.

This computer version of H differs from the analytical version in that there is a redundant argument $Z_3 \equiv Z_1 - Z_2$, included to improve numerical accuracy. It is desirable for a computer code to be able to deal with nearly any case allowed by the mathematical theory. One allowed case is that in which Z_1 and Z_2 are nearly equal in the sense that the difference $Z_1 - Z_2$ is a tiny fraction of either of the two Z's. If the difference is calculated by letting the computer subtract the nearly equal Z's, the relative error will be large unless the Z's are passed to H with sufficiently high precision and the subtraction performed with the same precision. Error is especially disruptive if the difference is close to the forbidden value -1 , because H is singular when $1 + Z_1 - Z_2 = 0$. An alternative to passing two high-precision arguments (and performing high-precision arithmetic) is to pass three lower precision arguments with the difference being one of the arguments. Of the three arguments, only the two having the smallest absolute values really need be passed. But the subprogram allows Z_1 and Z_2 to be any pair of numbers in the domain of H, so it is not known in advance which argument has the largest absolute value, and any one of the three can have the smallest absolute value. Therefore, all three arguments are passed.

A tolerance parameter DELTOL is set equal to 10^{-4} . This would result in $H(Z_1, Z_2, Z_3)$ being calculated with an error of less than one part per ten thousand (the intended accuracy), if machine precision was unlimited. An effort was made to manipulate expressions into forms that do not subtract nearly equal numbers. In spite of this effort, machine precision can still limit the accuracy in some extreme cases, such as when Z_3 is very close to -1 where H is undefined. The intended accuracy is not always guaranteed, and will still not be guaranteed even if DELTOL is assigned a smaller value.

FUNCTION H(Y1, Y2, Y3)

C This function subprogram can be appended to a FORTRAN source
C code, allowing the code to call the function H defined in the
C text. This computer version of H differs from the analytical
C version in that there is a redundant argument $Y3 = Y1 - Y2$. All
C three arguments are passed to insure that the two having the
C smallest absolute values are represented with the greatest
C possible numerical precision. Y1 and Y2 cannot have opposite
C signs, and Y3 cannot be -1 .

C

DELTOL=1.0E-4

C

C Check for illegal arguments.

C

```
IF (Y1*Y2.LT.0.0) THEN
  WRITE(*,*)'ERROR: FIRST TWO ARGUMENTS HAVE OPPOSITE SIGNS'
  GO TO 130
END IF
```

C

```
IF (Y3.EQ.-1.0) THEN
  WRITE(*,*)'ERROR: THIRD ARGUMENT IS -1'
  GO TO 130
END IF
```

C

C Assign a new value to the argument having the largest absolute
C value if needed to comply with $Y_3=Y_1-Y_2$, without disturbing
C the other two arguments.

C

```
IF (Y2*Y3.GE.0.0) Y1=Y2+Y3
IF (Y1*Y3.LE.0.0) Y2=Y1-Y3
```

C

C If Y_1 and/or Y_2 are negative, use $H(Y_1, Y_2, Y_3) = -H(-Y_2, -Y_1, Y_3)$.
C The arguments used will be Z_1 , Z_2 , and Z_3 . Set a flag as a
C reminder to multiply H by -1 if the Z 's differ from the Y 's.

C

```
IFLAG=0
Z1=Y1
Z2=Y2
Z3=Y3
IF ((Y1.LT.0.0).OR.(Y2.LT.0.0)) THEN
  Z1=-1.0*Y2
  Z2=-1.0*Y1
  IFLAG=1
END IF
```

C

C Use special cases or asymptotic forms if applicable. T with
C or without subscripts is for temporary storage of intermediate
C results and can represent different quantities in different
C calculations.

C

```
IF (Z3.EQ.0.0) THEN
  E=Z2
  GO TO 120
END IF
```

C

```
IF (Z2.EQ.0.0) THEN
  E=0.0
  GO TO 120
END IF
```

C

```
AZ3=ABS(Z3)
IF (AZ3.LT.DELTOL) THEN
  E=Z2
  GO TO 120
END IF
```

C

```
IF (Z3.LE.0.0) GO TO 10
```

```

T=Z3*(DELTOL+1.0)/(Z2*DELTOL)
T1=ALOG(1.0+T)
T2=1.0/T1
IF (Z2.LT.T2) THEN
    E=Z2
    GO TO 120
END IF
10 CONTINUE
C
AZD=ABS(1.0+Z3)/2.0
IF (AZD.LT.DELTOL) THEN
    E=0.5*(Z1+Z2)/(1.0+Z3)
    GO TO 120
END IF
C
IF (Z1.LE.0.0) GO TO 20
IF (Z2.LT.1.0) GO TO 20
T=(Z1+0.5)/(Z2-0.5)
T1=ALOG(T)
T2=2.0*DELTOL*(Z2-0.5)*T1/(1.0+Z3)
T3=1.0/T2
IF (Z1.GT.T3) THEN
    E=1.0/T1
    GO TO 120
END IF
20 CONTINUE
C
C If none of the above cases apply, prepare to estimate X2 by
C constructing a lower bound XA and an upper bound XB. Start
C with bounds that always apply and then go through the list to
C see whether the upper bound can be made smaller or the lower
C bound made larger.
C
XB=1.0
T=1.0+4.0*Z1*Z2
T1=SQRT(T)
T2=1.0/(1.0+T1+2.0*Z1)
XA=2.0*T2*(1.0+Z3)
C
IF (Z3.GT.-1.0) GO TO 30
IF (Z1.LE.0.0) GO TO 30
XAN=(1.0+Z3)/Z1
IF (XAN.GT.XA) XA=XAN
30 CONTINUE
C
XBN=(1.0+Z3)/(Z1+Z2)
IF (XBN.LT.XB) XB=XBN
C
IF (Z3.LT.0.0) GO TO 40
T=(1.0+Z3)/Z2
T1=T-ALOG(1.0+T)
XBN=Z2*T1/Z3
IF (XBN.LT.XB) XB=XBN
40 CONTINUE
C

```

```

IF (Z2.LT.1.0) GO TO 50
IF (Z3.GT.0.0) GO TO 50
T=(1.0+Z3)/(Z2-0.5)
T1=ALOG(1.0+T)
T2=Z2*T1/Z3
T3=(1.0+Z3)/Z3
XBN=T3-T2
IF (XBN.LT.XB) XB=XBN
50 CONTINUE
C
IF (Z2.LT.1.0) GO TO 60
IF (Z3.LT.0.0) GO TO 60
T=(1.0+Z3)/(Z2-0.5)
T1=ALOG(1.0+T)
T2=Z2*T1/Z3
T3=(1.0+Z3)/Z3
XAN=T3-T2
IF (XAN.GT.XA) XA=XAN
60 CONTINUE
C
IF (Z3.LT.-1.0) GO TO 70
IF (Z3.GT.0.0) GO TO 70
T=(1.0+Z3)/Z2
T1=T-ALOG(1.0+T)
XAN=Z2*T1/Z3
IF (XAN.GT.XA) XA=XAN
70 CONTINUE
C
IF (Z1.LE.0.0) GO TO 80
IF (Z3.GT.-1.0) GO TO 80
IF (Z2.LT.1.5) GO TO 80
T=(1.0+Z3)/(Z2-1.0)
T1=ALOG(1.0+T)
T2=Z2*T1/Z3
T3=(1.0+Z3)/Z3
XAN=T3-T2
IF (XAN.GT.XA) XA=XAN
80 CONTINUE
C
IF (Z3.LT.0.0) GO TO 90
T0=ALOG(1.0+Z3/Z2)
T0=1.0/T0
IF (Z2.GE.T0) GO TO 90
T=EXP(-1.0/Z2)
T1=(1.0-T)/(Z2-Z1*T)
T2=Z2*T1/Z3
T3=(1.0+Z3)/Z3
XAN=T3-T2
IF (XAN.GT.XA) XA=XAN
90 CONTINUE
C
C If XA and XB are so close together that numerical error gave
C  $XB \leq XA$ , use  $X2 = (1/2)(XA + XB)$  and calculate E and skip the
C bisection loop.
C

```

```

IF (XB.LE.XA) THEN
  X=0.5*(XA+XB)
  T=(1.0+Z3)-Z3*X
  E=Z2/T
  GO TO 120
END IF
C
C Now calculate WA and WB and record their signs in SA and SB.
C
X1=(Z1*XB-(1.0+Z3))/Z2
WB=1.0
IF (XB.LT.1.0) WB=X1-XB+ALOG((1.0-X1)/(1.0-XB))
SB=1.0
IF (WB.EQ.0.0) SB=0.0
IF (WB.LT.0.0) SB=-1.0
C
X1=(Z1*XA-(1.0+Z3))/Z2
WA=X1-XA+ALOG((1.0-X1)/(1.0-XA))
SA=1.0
IF (WA.EQ.0.0) SA=0.0
IF (WA.LT.0.0) SA=-1.0
C
C If XA or XB are so close to the correct solution X2 that
C numerical error gave  $XB \leq X2$  or  $XA \geq X2$ ,  $SA*SB$  will be positive
C or zero. If  $SA*SB = -1$ , everything is okay and the next block
C of steps can be skipped. Otherwise, determine which of the
C intended bounds is closest to  $X2$ . Set  $X2$  equal to that
C intended bound and calculate  $E$  and skip the bisection loop.
C
IF (SA*SB.EQ.-1.0) GO TO 100
T=(1.0+Z3)-Z3*XA
IF (SB*WB.LT.SA*WA) T=(1.0+Z3)-Z3*XB
E=Z2/T
GO TO 120
100 CONTINUE
C
C Now start the bisection loop to tighten the  $X2$  bounds.
C
110 CONTINUE
X=0.5*(XA+XB)
X1=(Z1*X-(1.0+Z3))/Z2
W=X1-X+ALOG((1.0-X1)/(1.0-X))
S=1.0
IF (W.LT.0.0) S=-1.0
IF (S*SB.GT.0.0) XB=X
IF (S*SB.LT.0.0) XA=X
C
C If  $W=0$ , the solution was found. Calculate  $E$  and exit from the
C loop.
C
IF (W.EQ.0.0) THEN
  T=(1.0+Z3)-Z3*X
  E=Z2/T
  GO TO 120
END IF

```

C
C EA and EB are the values of E when X2 is replaced with XA and
C XB respectively. If EA and EB are close enough together,
C calculate E and exit from the loop. Otherwise, go through the
C loop again.
C

```
T=(1.0+Z3)-Z3*XA
EA=Z2/T
T=(1.0+Z3)-Z3*XB
EB=Z2/T
E=2.0*EA*EB/(EA+EB)
T=(EA-EB)/(EA+EB)
DELTA=ABS(T)
IF (DELTA.LT.DELTOL) GO TO 120
GO TO 110
120 CONTINUE
H=E
IF (IFLAG.EQ.1) H=-1.0*E
130 CONTINUE
RETURN
END
```

APPENDIX C: THE SPECIAL FUNCTION F

The function F is defined by $Y=F(X_1, X_2)$ if and only if Y satisfies

$$Y + (1 - X_1) \ln(1 + Y/X_1) = X_2 . \quad (C1)$$

It is sufficiently general to confine our attention to those cases where X_1 is positive and X_2 is positive or zero. X_1 is not allowed to be zero.

The function F is closely related to a particular type of inverse of the special function H. If we want to solve (C1) for X_1 when Y and X_2 are given, the solution can be expressed in terms of H. If we want to solve (C1) for Y with the X's given, we use F. But F is much simpler than H and an approximation for F was already listed in the form of the generalized ambipolar approximation. The connection is made clear by remembering where F first originated. For the p-type substrate with $g=0$, Section 3.2 found that

$$P + (p_0/2 - A) \ln(1 + P/A) = \Omega$$

or

$$P = (p_0/2) F(2A/p_0, 2\Omega/p_0) . \quad (C2)$$

The generalized ambipolar approximation is an approximation for either side of (C2).

Iterations are used to evaluate F, or solve for Y. Iterations are performed by manipulating (C1) into

$$Y = f(X_1, X_2, Y) \quad (C3)$$

for some appropriately chosen f, which is not unique. If f is chosen well, the sequence of iterates $Y^{(0)}, Y^{(1)}, \dots$ will con-

verge to the solution, where $y^{(0)}$ is some initial guess and

$$y^{(i+1)} \equiv f(X_1, X_2, y^{(i)}) \quad (i = 0, 1, \dots) .$$

Several cases, characterized by the way X_1 and X_2 compare with each other, are considered separately. The need for considering different cases can be seen from the fact that P given by (C2) behaves differently under different conditions. If $A \geq p_0/2$ (implying that $V_2 \leq 0$), the nominal ambipolar approximation is a good low-order approximation. If $A < p_0/2$ (implying that $V_2 > 0$), there is an HRR and an AR, and the behavior of P depends upon which region we are examining. We can anticipate that at least three cases require separate treatment. It turns out that there are four cases, with one corresponding to a transitional region near the ARB.

The different cases will use different f 's in (C3) and different intervals from which the initial guess is to be selected. The proof of convergence is fundamentally the same for all cases. The basic idea is to find a closed interval, from which the initial guess is to be selected, having the property that f , regarded as a function of Y , maps this interval into itself. Then show that, throughout this interval, the absolute value of the Y derivative of f is less than or equal to some number that is strictly less than 1 (preferably less than $1/2$ so that the iteration will converge at least as fast as the bisection method). The details are omitted because they are not difficult. Error estimates are also obtained by iteration, but not necessarily the same convergent iteration that produces progressively better estimates. The basic idea is the same for all cases and illustrated for the first case considered.

We start with Case 1 defined by

$$X_1 \geq 1 \quad (\text{defines Case 1}) .$$

This case will be encountered when we want to use (C2) to solve for P and there is no HRR (i.e., $A \geq p_0/2$ or $V_2 \leq 0$). The iteration is

$$Y^{(i+1)} = X_2 + (X_1 - 1) \ln(1 + Y^{(i)}/X_1) \quad (\text{for Case 1}) \quad (C4)$$

which converges for any initial guess selected from the interval $[X_2, \infty)$. The suggested initial guess is

$$Y^{(0)} = X_2 \quad (\text{for Case 1}) .$$

Convergence of the iteration (C4) can be very slow in theory. In practice, Case 1 is accompanied by $X_1 \approx 1$ and the convergence is fast. An error estimate associated with any given iterate is obtained by manipulating (C1) into

$$Y = Y + [(X_2 + X_1)/(X_2 + 1)] [X_2 + (X_1 - 1) \ln(1 + Y/X_1) - Y] .$$

The actual solution Y and all iterates produced by (C4) are in the interval $[X_2, \infty)$. Differentiating shows that the right side of the above equation is decreasing in Y on this interval (or constant if $X_1=1$). Therefore the right side maps iterates that are too small into estimates that are too large and vice-versa. The correct solution is bracketed by any iterate $Y^{(i)}$ and its conjugate $Y_C^{(i)}$ defined by

$$Y_C^{(i)} \equiv Y^{(i)}$$

$$+ [(X_2+X_1)/(X_2+1)] [X_2+(X_1-1)\ln(1+Y^{(i)}/X_1)-Y^{(i)}] \quad (\text{for Case 1}) .$$

The difference between $Y^{(i)}$ and $Y_C^{(i)}$ is a simple error estimate.

Case 2 is defined by

$$0 < X_1 < 1 \quad \text{and}$$

$$X_2 > 2 + (1 - X_1) \ln(1 + X_2/X_1) \quad (\text{defines Case 2}) .$$

It can be shown that Case 2 implies that $Y > 2$, or $P > p_0$ in (C2).

This case is encountered when there is an AR and HRR ($A < p_0/2$) and we want to use (C2) to solve for P at some point in the AR not too close to the ARB. The iteration is

$$y^{(i+1)} = y^{(i)} + [(X_2 + X_1)/(X_2 + 1)] [X_2 + (X_1 - 1) \ln(1 + y^{(i)}/X_1) - y^{(i)}] \quad (\text{for Case 2})$$

which converges for any initial guess selected from the interval $[2, X_2]$. The suggested initial guess is

$$y^{(0)} = X_2 \quad (\text{for Case 2}) .$$

The conjugate of a given iterate is either the next or previous iterate, i.e., the solution is bracketed between any pair of adjacent iterates.

Case 3 is defined by

$$0 < X_1 < 1 \text{ and}$$

$$(1-X_1)\ln[(1-X_1)/X_1] < X_2 \leq 2 + (1-X_1)\ln(1+X_2/X_1) \quad (\text{defines Case 3})$$

and is encountered when there is an HRR and we want to use (C2) to solve for P at a point close to and on either side of the ARB (a transitional region). The iteration is

$$y^{(i+1)} = y^{(i)} - (1/2) [y^{(i)} + (1 - X_1) \ln(1 + y^{(i)}/X_1) - X_2] \quad (\text{for Case 3})$$

which converges for any initial guess selected from the interval $[1/2 - X_1, X_2]$. Note that $1/2 - X_1$ can be the initial guess even when negative, but it is not a very good guess when negative. The suggested initial guess is

$$y^{(0)} = x_2 \quad (\text{for Case 3}) .$$

The solution is bracketed by any iterate $y^{(i)}$ and its conjugate $y_C^{(i)}$ defined by

$$y_C^{(i)} \equiv x_2 - (1 - x_1) \ln(1 + y^{(i)}/x_1) \quad (\text{for Case 3}) .$$

Note that the iteration can be written more concisely as

$$y^{(i+1)} = (1/2) [y^{(i)} + y_C^{(i)}] \quad (\text{for Case 3}) .$$

Case 4 is defined by

$$0 < x_1 < 1 \quad \text{and}$$

$$0 \leq x_2 \leq (1 - x_1) \ln[(1 - x_1)/x_1] \quad (\text{defines Case 4}) .$$

This case will be encountered when there is a wide HRR and we want to use (C2) to solve for P at some point in the HRR not too close to the ARB. The iteration is

$$y^{(i+1)} = (1/2) [y^{(i)} - x_1] \\ + (x_1/2) \cdot \exp[(x_2 - y^{(i)})/(1 - x_1)] \quad (\text{for Case 4})$$

which converges for any initial guess selected from the interval $[0,1]$. The suggested initial guess is

$$y^{(0)} = 0 \quad (\text{for Case 4}) .$$

The solution is bracketed by any iterate $y^{(i)}$ and its conjugate $y_C^{(i)}$ defined by

$$Y_C^{(i)} \equiv X_1 \exp[(X_2 - Y^{(i)})/(1 - X_1)] - X_1 \quad (\text{for Case 4}) .$$

Note that the iteration can be written more concisely as

$$Y^{(i+1)} = (1/2) [Y^{(i)} + Y_C^{(i)}] \quad (\text{for Case 4}) .$$

The following function subprogram can be appended to a FORTRAN source code, allowing the code to call the function F as it would call any built-in function. The iterations are terminated when error estimates indicate that the sum $F(X_1, X_2) + X_1$ has an error less than one part per ten thousand. The number of iterations needed to produce this accuracy depends on the individual case. The number can be as large as twelve or thirteen (comparable to the bisection method) or as small as two or three.

```

FUNCTION F(X1,X2)
C This function subprogram can be appended to a FORTRAN source
C code, allowing the code to call the function F defined in the
C text. X1 must be positive and X2 must be nonnegative.
C
      DELTOL=1.0E-4
C
C Check for illegal arguments.
C
      IF (X1.LE.0.0) THEN
        WRITE(*,*) 'ERROR: X1 IS NOT POSITIVE'
        GO TO 100
      END IF
      IF (X2.LT.0.0) THEN
        WRITE(*,*) 'ERROR: X2 IS NEGATIVE'
        GO TO 100
      END IF
C
C Determine which of the four cases apply and go to the
C appropriate block.
C

```

```
IF (X1.GE.1.0) GO TO 10
XH=2.0+(1.0-X1)*ALOG(1.0+X2/X1)
XL=(1.0-X1)*ALOG((1.0-X1)/X1)
IF (X2.GT.XH) GO TO 30
IF ((X2.LE.XH).AND.(X2.GT.XL)) GO TO 50
GO TO 70
```

C

C Case 1 block starts here.

C

```
10 CONTINUE
Y=X2
20 CONTINUE
Y=X2+(X1-1.0)*ALOG(1.0+Y/X1)
T=X2+(X1-1.0)*ALOG(1.0+Y/X1)-Y
YC=Y+(X2+X1)*T/(X2+1.0)
ERROR=ABS(Y-YC)/(Y+X1)
IF (ERROR.GT.DELTOL) GO TO 20
GO TO 90
```

C

C Case 2 block starts here.

C

```
30 CONTINUE
Y=X2
40 CONTINUE
YC=Y
T=X2+(X1-1.0)*ALOG(1.0+Y/X1)-Y
Y=Y+(X2+X1)*T/(X2+1.0)
ERROR=ABS(Y-YC)/(Y+X1)
IF (ERROR.GT.DELTOL) GO TO 40
GO TO 90
```

C

C Case 3 block starts here.

C

```
50 CONTINUE
Y=X2
YC=X2-(1.0-X1)*ALOG(1.0+Y/X1)
60 CONTINUE
Y=0.5*(Y+YC)
YC=X2-(1.0-X1)*ALOG(1.0+Y/X1)
ERROR=ABS(Y-YC)/(Y+X1)
IF (ERROR.GT.DELTOL) GO TO 60
GO TO 90
```

C

C Case 4 block starts here.

C

```
70 CONTINUE
   Y=0.0
   T=(X2-Y)/(1.0-X1)
   YC=X1*EXP(T)-X1
80 CONTINUE
   Y=0.5*(Y+YC)
   T=(X2-Y)/(1.0-X1)
   YC=X1*EXP(T)-X1
   ERROR=ABS(Y-YC)/(Y+X1)
   IF (ERROR.GT.DELTOL) GO TO 80
90 CONTINUE
   F=Y
100 CONTINUE
   RETURN
   END
```

REFERENCES

- [1] C.M. Hsieh, P.C. Murley, and R.R. O'Brien, "A Field-Funneling Effect on the Collection of Alpha-Particle-Generated Carriers in Silicon Devices," IEEE Electron Device Letters, Vol.EDL-2, No.4, pp.103-105, April 1981.
- [2] L.D. Edmonds, A Generalized Law of the Junction for p-n Junctions in Silicon Devices, Jet Propulsion Laboratory Publication 92-24, October 1, 1992.
- [3] J.R.A. Beale and J.A.G. Slatter, "The Equivalent Circuit of a Transistor with a Lightly Doped Collector Operating in Saturation," Solid-State Electronics, Vol.11, pp.241-252, 1968.
- [4] P.M. Morse and H. Feshbach, Methods of Theoretical Physics, McGraw-Hill, p.24, 1953.
- [5] M.R. Pinto, C.S. Rafferty, H.R. Yeager, and R.W. Dutton, PISCES-IIB Supplementary Report, Stanford Electronics Laboratories, Department of Electrical Engineering, Stanford University, Stanford CA 94305, 1985. (Modified to include cylindrical coordinates and photogeneration.)
- [6] A. Herlet, "The Forward Characteristic of Silicon Power Rectifiers at High Current Densities," Solid-State Electronics, Vol.11, p.717, 1968.