INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106

TECHNIQUES FOR MODELING MULTIDIMENSIONAL EFFECTS IN INTEGRATED DEVICES

by

Jakob Hans Hohl

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF ELECTRICAL ENGINEERING

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

1 9 7 6

THE UNIVERSITY OF ARIZONA

GRADUATE COLLEGE

I hereby recommend that this dissertation prepared under my direction by Jakob Hans Hohl

entitled Techniques for Modeling Multidimensional Effects in Integrated Devices

be accepted as fulfilling the dissertation requirement of the degree of Doctor of Philosophy

Douglas J Hamilton — April 13, 1976
Dissertation Director — Date

After inspection of the final copy of the dissertation, the following members of the Final Examination Committee concur in its approval and recommend its acceptance:*

R. L. Gallo	14 Apr 1976
D J Hamilton	13 Apr 1976
John A. Reagan	14 Apr. 1976

*This approval and acceptance is contingent on the candidate's adequate performance and defense of this dissertation at the final oral examination. The inclusion of this sheet bound into the library copy of the dissertation is evidence of satisfactory performance at the final examination.

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Yakob Hans Helm

To Annemarie

ACKNOWLEDGMENTS

I wish to express my sincere gratitude to Dr. Douglas J. Hamilton, my advisor, for his support and encouragement throughout this doctoral program.

Many stimulating discussions with members of the faculty and staff of the Department of Electrical Engineering and the Department of Physics of The University of Arizona are also gratefully acknowledged. They contributed to the clarification of ideas crucial to this work.

I also wish to thank Dr. Stanley Rush of the Department of Electrical Engineering of The University of Vermont for assuming responsibility as local advisor, and for helpful suggestions.

The successful completion of the study program was greatly facilitated by a two year Resident Study Fellowship from the IBM Corporation.

Last, but not least, I should like to thank Mrs. Therese Clark for her expertise and care in typing this difficult manuscript.

TABLE OF CONTENTS

TABLE OF CONTENTS--Continued

LIST OF ILLUSTRATIONS

LIST OF TABLES

ABSTRACT

This work is devoted to the development and evaluation of techniques for calculating impurity profiles, carrier distributions, and carrier flux distributions in semiconductor devices of contemporary integrated circuits. While two- and three-dimensional formulations are the main objectives, improved methods for calculating one-dimensional distributions are also included.

Successful techniques for modeling impurity profiles are analytic or semi-analytic in nature, while strictly numerical algorithms prove too inaccurate or too uneconomical. The accuracy of the newly developed techniques is shown to be compatible with modern semiconductor fabrication methods.

In order ultimately to obtain semiconductor device models of consistent overall accuracy and detail the effects of high impurity concentrations have been included in the analytic description of the carrier statistics. This has been done without introducing major complications into the basically simple formulation.

The majority carrier concentrations in nonuniformly doped semiconductor regions away from the metallurgical junctions, metal contacts and surfaces are shown to be almost exactly congruent with the net impurity concentrations. The carrier distributions in the remaining regions, and in nonequilibrium states cannot be described adequately by analytic formulations. These problems must be solved numerically.

The only promising technique for this purpose seems to be the finite element method, which is based on a variational principle. The equilibrium state of a semiconductor region is characterized by minimum Helmholtz free energy. The corresponding functional is derived and shown to be less trivial than might have been expected.

It is also demonstrated that in weak nonequilibrium states the power dissipation is minimized. This formulation clearly points out that the technically important operating conditions of devices are beyond the limits of validity of this variational principle.

CHAPTER 1

INTRODUCTION

Many engineers and scientists who have been developing techniques for computer-aided circuit design and analysis have, for a long time, harbored the dream of composing a fully automatic circuit design and analysis system. Such a computer program would be capable of accepting the description of an integrated circuit in the form of the physical parameters of the semiconductor processing steps, and of the numerical data describing the photomask layouts. From this information it would automatically construct electrical equivalent circuit models, and from these it would calculate responses to applied electrical signals, without further manual interventions. This ambitious goal has, so far, not been reached in spite of strong efforts by many outstanding teams, although great strides have been made toward it.

The main steps in such an automatic circuit design system would be,

1. The determination of the topology of the integrated circuit, and of the layout geometries of the individual devices from the photomask information,
2. The computation of the impurity profiles for each device from its layout and from the physical process parameters,

3. The determination of the electrical terminal characteristics, i.e., of the electrical model, of each device from the impurity profiles,
4. The assembly of the electrical model of the integrated circuit from the device models and from the circuit topology.
5. The calculation of the response of the circuit model to applied signals, i.e., the simulation of the electrical behavior of the integrated circuit.

The first and the last two steps have been developed to the most sophistication and can be considered solved. Computer programs which perform step (1) are incorporated in computer-aided circuit layout and mask generating systems. Such programs are closely tailored to the specialized data processing equipment used in a particular installation. They would not be suitable for general applications outside these systems and have not become generally available. Without such automatic mask generating systems the integrated circuit technology could not have evolved to the present state of the art of large scale integration.

The last two steps, (4) and (5), are performed by many well established and generally available circuit analysis programs, such as CORNAP [1], or ASTAP [2], to mention but two.

The development of the techniques to perform steps (2) and (3) have been lagging behind the advances of the integrated circuit technology to various degrees. The original device analysis and modeling techniques have evolved mostly in parallel with the early

development of discrete semiconductor devices. These device structures had geometries which can be modeled adequately by one-dimensional formulations. Furthermore, the variability of the processing procedures was of a magnitude that even analyses which entailed crude and simple analytic approximations still yielded results of sufficient accuracy. In spite of their limitations, these simple approximations provide excellent insight into the essential characteristics of the processing steps and of the phenomena which determine the electrical behavior of the devices. The simplicity and conceptual transparence of these analysis techniques have made them the foundation of all basic education in semiconductor device technology. The resulting simple device models, as, for example, the well-known Ebers-Moll model, have found widespread application in circuit analysis and simulation.

Although these conventional device models represent an anachronism compared with the sophistication of the modern integrated circuit processing methods and network analysis techniques, they are still widely used, not in the least because of the lack of improved models which exhibit a good balance between accuracy, simplicity and versatility. Hence, the device models now form the weakest link in the simulation of contemporary integrated circuits, and they limit the overall accuracy of the analysis. At the same time, circuit analysis and simulation become more and more indispensable for the development of integrated circuits. While the analysis and optimization of circuits with discrete components can be performed swiftly by a combination of calculations and measurements, the design of integrated circuits is in many cases completely dependent on simulation. Many of

the nodes of such circuits are buried underneath layers of insulators and conductors, and are therefore inaccessible to probing. Other nodes may be accessible in principle, but they are too small to be contacted reliably by a probe tip. But above these mechanical limitations, the majority of integrated circuits operates at such low energy levels that the disturbances introduced by probes would alter the performance to the extent that the measurements would become useless. If, in these cases, the analysis also produces inaccurate results, the designs will be less than optimum, which eventually can lead to reduced manufacturing yield.

A similar situation exists with respect to the device models themselves. The early, simple models could easily be compared with actual devices by measurement, as long as the device structures were large. With the progressing miniaturization of the devices the measurements become more and more unreliable because they interfere with the device performance, or because of the limitations of the instruments. In these cases only the more accurate mathematical modeling of the small devices remains as an alternative. Reliable measurements can still be performed on larger or specially designed test structures. Such measurements determine overall electrical or process parameters.

Of course, the device analysis efforts did not cease with the inception of the early models. These activities proceeded mainly along two lines. On the one side, innumerable specific effects and aspects of the various devices have been analyzed and described in great detail by analytic techniques. These investigations were usually done in one-dimensional formulation because they would be intractable otherwise.

The results of these specialized and disjointed developments are not well suited as building blocks of a general device analysis system.

Another line of development, more oriented toward versatility, has relied exclusively on numerical techniques. Such a development was natural in view of the rapid evolution of computer systems which has gone hand in hand with the advances of the integrated circuit technology. The strictly numerical modeling mainly consists of the simulation of the partial differential equations describing impurity diffusion and carrier transport, using finite difference algorithms. These methods would in general be versatile enough to simulate many different device structures and would not be restricted to one-dimensional formulation. However, in applying them one not only trades the difficulties of developing more refined analytic techniques for the grief of debugging elaborate computer programs and refining numerical algorithms, but also the insight provided by the analytic formulations is lost in the reams of numbers which represent the numerical solutions.

The absence of numerical solutions of three-dimensional diffusion or carrier transport problems is noteworthy. It has mainly two reasons. The finite difference techniques would require tens of thousands of nodes just to achieve numerical stability, let alone accuracy. This not only exceeds the capacity of most computers of reasonable size, but the computing times become excessive and economically unbearable. Further, the pictorial representation of three-dimensional solutions is very cumbersome. On the whole, the strictly numerical methods have not nearly produced what they may have

promised initially. This work will shed some light on their inherent limitations.

In spite of the sophistication of the modern circuit analysis programs the efficient simulation of more complex circuits still requires comparatively simple device models which can be described with few parameters or simple functions. Because of this, a number of contributions [3, 4, 5] have been specifically directed toward methods for deriving simple electrical device models, the complexity of which could be adapted to the requirements of the circuit simulation. These activities have culminated in the formulation of a systematic method for deriving electrical device models for one-dimensional structures by Hamilton, Lindholm, and Marshak [6]. Fossum [7] has extended it to two and three dimensions. The innovation of this method is a systematic technique of deriving an equivalent network of a device from the distributions of the carriers and potentials in its structure. The technique is based on the representation of the continuity equations for the carriers in a form analogous to the Kirchhoff equations of an electrical network. These equations are appropriately called the distributed Kirchhoff equations. When they are integrated over each of a number of judiciously chosen subregions of the device structure, the individual terms yield the component values of an equivalent circuit which has a node for every subregion.

This modeling method is particularly suitable for numerical techniques; it provides an algorithm with which the reams of numbers constituting the numerical solution of the carrier transport problem can be reduced to the few parameters of an equivalent network

representation of the device. Nevertheless, the method has not yet found wide application in computer-aided analysis of integrated circuits. The main reason may lie in the difficulties with calculations of impurity profiles and of distributions of carriers, carrier fluxes, and potentials.

This study is aimed at improving the methods for calculating these distributions. In view of the severe limitations of strictly numerical calculations, the efforts are primarily directed at refining analytic approximations. This is done under careful observation of a good balance between accuracy and simplicity of the formulations. The numerical calculations are relegated to the eventual evaluation of the analytic expressions.

Chapter 2 is devoted to the calculation of one-, two-, and three-dimensional impurity profiles constituting the device structures. Important profiles which hitherto were either determined by numerical methods, or not at all, can be described with sufficient accuracy by simple analytic approximations. The finite difference method and the finite element method, on the other hand, prove too inaccurate for calculating impurity profiles, and only a semi-analytic approach based on Green's functions yields three-dimensional profiles of proven accuracy.

Chapters 3 and 4 describe extensions of the conventional semiconductor theory and related topics. The great success of the early models strongly suggests that strides be made to incorporate extensions in a way which leaves the basic structure of this theory intact.

The two main topics of Chapter 3 are the inclusion of high concentration effects into the analytic theory and the description of carrier distributions in nonuniformly doped semiconductor regions. These extensions of the theory are made under the assumption of thermostatic equilibrium, because only under this condition are the carrier distributions solely dependent on the impurity profiles. The incorporation of the high concentration effects leads only to minor complications of the analytic formulation. The local carrier concentrations in nonuniformly doped regions turn out to be related to the local impurity concentration in the same way as in homogeneous material, provided that the regions are distinctly extrinsic and not immediately adjacent to surfaces, metal contacts, or metallurgical junctions.

Chapter 3 also contains a review of the thermodynamic theory of carrier behavior and a survey of the phenomena occurring at the three types of interfaces just mentioned. The latter investigations fathom the limits of analytic approaches to device modeling. The strong nonlinearities in the equations which govern the behavior of carriers prohibit the application of superposition techniques, and quickly make the problems analytically intractable. Thus, numerical methods finally become unavoidable for the investigation of carrier transport phenomena, especially in three-dimensional situations. Even the determination of equilibrium concentrations in the vicinity of surfaces, contacts and metallurgical junctions calls for numerical solutions, and the simulation of the differential equations themselves becomes inevitable. The finite element method appears more promising than the finite difference

method, because it is not restricted to a regular lattice of nodes and is usually more stable numerically. It requires, however, a variational formulation of the phenomenon under investigation.

Chapter 4 is concerned with variational principles governing the carrier behavior at equilibrium and weak nonequilibrium. The use of the finite element method without a thoroughly established variational principle would be very risky, because an incorrect functional would inevitably lead to wrong solutions, without making the algorithms inoperative. The validity and correctness of a variational formulation should be proven analytically by deriving the correct differential equations by varying the functional. The two variational formulations discussed in Chapter 4 pass this test.

Since the attributes of integrated devices investigated here cannot be measured directly to any reasonable degree of accuracy, no attempt has been made to verify the results experimentally, and consequently no experimental data are reported in this dissertation. Instead, considerable efforts have been devoted to the determination or estimation of the errors of the proposed methods and approximations.

CHAPTER 2

PROBLEMS OF IMPURITY DIFFUSION

2.1 Outline of the Chapter

Impurity diffusion through windows in the masking layer results in three-dimensional impurity profiles. As long as the window dimensions are large in comparison to the depth of the diffusion, a one-dimensional description will approximate the diffusion profile sufficiently well over most of the area of the window, and the edge effects are in many cases of little consequence. Grove [8, Chapter 3] gives a comprehensive outline of the techniques, analyses, problems, and results of impurity diffusion processes used in the fabrication of integrated devices. The linear and one-dimensional predeposition problem has a convenient closed form solution. The impurity profile obtained after a subsequent drive-in diffusion has traditionally been approximated by another closed form solution, provided that the predeposition diffusion depth is much less than the drive-in diffusion depth. These approximations, although not reflecting the pronounced nonlinearities of the diffusion process at the high concentrations encountered during predeposition, have found wide use in the modeling of integrated devices and have proven remarkably successful on the whole. Their value probably lies in the closed form descriptions of the profiles rather than in the accuracy.

With the progressing miniaturization of integrated devices, the diffusion window dimensions have become comparable to the diffusion depth, and also, the predeposition diffusion depth has in many instances become a significant fraction of the drive-in diffusion depth. Unfortunately neither the one-dimensional two step diffusion profile nor the three dimensional profile generated by the diffusion through a small rectangular window--the shape most widely used--have closed form descriptions.

In view of the undisputed success of the conventional approximations mentioned above, it seems advantageous to pursue approximate closed form solutions more than exact numerical calculations of impurity profiles, and this is the guiding philosophy of this chapter. Although the numerical calculation of three dimensional profiles will figure prominently, it is not an end in itself; the solutions obtained are to serve only as standards against which the quality of closed form approximations can be measured.

The next section contains a short review of the technological aspects of integrated device fabrication, tailored to specifying the various diffusion problems and the appropriate boundary conditions.

The third section is devoted to one-dimensional solutions and approximations. It contains new approximations which quite accurately describe profiles from linear two step diffusions. One-dimensional solutions are important in their own right because even highly miniaturized structures contain many large diffusion areas. The one-dimensional problems also represent an excellent proving ground for tackling the more complicated multidimensional problems.

The fourth section, the longest of the chapter, concerns itself with the difficulties of finding accurate numerical solutions for the three-dimensional problem of predeposition through a rectangular window. The most straightforward numerical method, the finite difference method, which models the differential quotients by ratios of finite differences, turns out to be beyond the capabilities of computers of reasonable size because of the large number of lattice points required to assure an adequate description of the profiles. The next formal numerical method in line, the finite element method, rests on the variational formulation of the diffusion problem. Although this technique proves unsuccessful in the end also, it is worth dwelling on somewhat because of some pitfalls contained in the variational formulation of the time dependent diffusion problems, which are not addressed in many of the books on the subject [9, 10, 11]. The method which eventually produces acceptable numerical results is more heuristic than formal. Starting with the integral formulation of the boundary value problem and knowing that the only impurity sources are located in the window, one can try to adjust these sources in such a way that the boundary conditions are satisfied by the resulting impurity profile. To implement this technique strictly numerically would involve integration of the source distribution and Green's functions over the two spatial coordinates of the window and over the time, a time-consuming procedure of limited accuracy and insight. It turns out, however, that the time integrals of the Green's functions for sources with time dependence $t^{n/2}$, n = ..., -1, 0, 1, .., are expressible in closed form. A distribution of point sources in the

window with only time dependence $t^{-\frac{1}{2}}$ and t produces a profile which, for the time ranges of interest, satisfies the boundary conditions within a few per cent over most of the area of the window, except in a narrow strip along the edge. The profile exactly satisfies the boundary condition outside the window. Once a satisfactory set of sources is determined, the impurity concentration at any point in the semiconductor at any time during the predeposition diffusion can be calculated by merely adding the contributions of all the sources. This process takes less computing time than an iterative method and yields a result of the same relative accuracy as that of the sources, whether the local impurity concentration is of the order of the surface concentration in the window or many orders of magnitude lower. Herein lies the advantage of this trial and error procedure over the two other, formally more pleasing, methods. Those methods propagate errors over the entire lattice of points of support, and the results become meaningless in areas of low concentration. But the low-concentration regions are usually of more interest than those of high concentration.

The fifth section is devoted to analytical approximations of the three-dimensional impurity profile of a predeposition diffusion through a rectangular window. Gajda and Jackson [12] consider the window as a sector of the infinite plane and use a source strength which is constant over the area of the window and impulsive in the time. This approximation leads to a vertical, one-dimensional Gaussian profile far inside the window. A similar approximation, proposed as an example, leads to the correct, complementary error

function profile far inside the window. Displays of constant concentration curves illustrate the strengths and limitations of this particular approximation, and a method for improvement of the profile is suggested.

The final section of the chapter addresses the determination of predeposition profiles from diffusions with concentration dependent diffusivity. A transformation which relates the nonlinear diffusion equation with its linear counterpart allows the determination of nonlinear profiles in terms of the linear profiles of the same one-, two-, or three-dimensional geometries.

2.2 Technological Considerations

The successful fabrication of integrated circuits rests above all on the reproducibility of two technological processes: photolithography and impurity atom diffusion in the semiconductor crystal at high temperature. This chapter only considers the aspects of diffusion processes, with the assumption that the photolithographic steps are producing diffusion windows of perfect shape. Although the considerations refer to silicon, they are in principle equally applicable to other semiconductor materials.

In the fabrication of silicon integrated circuits, silicon dioxide is almost exclusively used as diffusion barrier. It can be grown conveniently by heating the silicon surface to above 800°C in an oxidizing atmosphere. Other barrier materials can be laid down by chemical vapor deposition. The raw silicon wafer is always doped with a surplus of either acceptor or donor atoms at background concentrations

of 10^{15} cm^{-3}, more or less. When oxide grows on such a wafer, not only silicon atoms are consumed, but also impurity atoms are built into the oxide which is known to be growing at the oxide-silicon interface. The doping level of the oxide is determined by the segregation coefficient which is the ratio of the impurity concentration in the oxide to that in the silicon next to the interface, in thermal equilibrium. Any oxidation will lead either to an accumulation or a depletion of impurities near the silicon surface. The diffusivity of impurities in the oxide is at least an order of magnitude lower than that in the silicon so that even after prolonged heating to high temperatures a negligible amount of impurities will escape from the silicon into a stationary oxide. Thus, considering the oxide-silicon interface at the start of a diffusion process without oxide growth as the xy-plane, with the z-axis pointing into the silicon, the mathematical boundary condition is that of zero flux of impurities through the xy-plane.

During diffusion underneath a growing oxide the silicon surface moves in the z-direction, and a flux of impurity atoms through this surface is the appropriate boundary condition. This flux is a function of the instantaneous impurity concentration in the silicon near the interface.

During the predeposition process, impurities diffuse from the doped atmosphere in the diffusion furnace through the windows in the barrier into the silicon. If the impurity concentration in the gas is low, the diffusion rate depends on the transport processes in the gas and hence on the ill-controllable flow dynamics of the gas around the wafers. When the gas is highly doped, the impurity concentration at the

silicon surface reaches the solid solubility and is no longer dependent on the gas flow characteristics. The solid solubility depends only on the species of impurity and is a slight function of the temperature in the range of interest. Solubility limited diffusion is therefore well controlled and is the standard process used for predeposition diffusion in integrated circuit fabrication. The resulting boundary condition in the window is a constant surface concentration, C_s, which assumes values of the order of 10^{20} cm^{-3}. The steep concentration gradients and high surface concentrations of the predeposition profile are often undesirable and a subsequent drive-in diffusion in an atmosphere devoid of impurities is carried out. Often the atmosphere is oxidizing for growing an oxide layer in the windows for further use in photolithographic and diffusion steps. If no oxide is grown, the boundary condition in the window is that of outdiffusion of impurities through the silicon surface. When an oxide is grown during the drive-in process, the impurities are almost instantaneously sealed off from the gas, and the boundary condition of the growing oxide discussed above prevails.

Since the impurity concentrations at the silicon surface reach as high as 10^{21} cm^{-3}, while the concentrations at the junctions are typically of the order of 10^{15} cm^{-3}, the solutions of the diffusion problems have to be reasonably accurate over five to six orders of magnitude of the concentration. This requirement imposes extreme demands on numerical solutions.

The mathematical treatment of the diffusion processes is often carried out under the assumption of linear diffusion with constant

diffusivity. Considerable deviations from these idealized conditions occur in practice. For one, the impurities ionize as they enter the silicon. As long as their concentration is much below that of the free carriers, any field effects are screened out by the much more mobile carriers. The high concentrations during the predeposition process, however, exceed that of the free carriers; and the resulting electric fields aid the dispersion of the impurities and increase the effective diffusivity. This effect is still essentially local. Nonlocal effects appear, for example, because of the lattice strains introduced by the high impurity concentrations. The stresses due to these deformations extend over large regions and change the effective diffusivities also. It is believed, for instance, that the effect known as "emitter push" is caused by stresses. Local effects can in principle be handled for geometries of high symmetry, provided that one knows the relation between the concentration and the diffusivity. Nonlocal effects are not only exceedingly hard to treat, but also little investigated and understood. They will not be addressed here.

2.3 Large Diffusion Windows--One-Dimensional Impurity Profiles

The diffusion process is governed by the conservation of impurities, expressed by the continuity equation

$$\nabla \cdot \vec{F} = -\partial C/\partial t, \tag{2.1}$$

and by Fick's law

$$\vec{F} = -D\nabla C. \tag{2.2}$$

C denotes the impurity atom concentration which will be measured in cm^{-3}; F stands for the flux of impurity atoms ($cm^{-2}h^{-1}$); and D represents the diffusivity ($cm^{2}h^{-1}$). The diffusivity is, in general, a second rank tensor, but the cubic symmetry of the semiconductor crystals reduces it to a scalar. This is true of all first order transport coefficients. Consider D as a function of C, expressed as

$$D = D(C) = D_o f(C). \tag{2.3}$$

Equation (2.2) is in this case really the definition of the diffusivity, and this interpretation is valid as long as the diffusion process is strictly a local process.

Insertion of (2.2) into (2.3) leads to the diffusion equation

$$\nabla\cdot(D\nabla C) = \partial C/\partial t. \tag{2.4}$$

It is worth noting that in all cases treated here, all impurities enter the semiconductor through its surface and none are generated in its interior. (Impurity deposition by ion implantation would be an example of internal sources.) Therefore the homogeneous Equation (2.4) holds throughout the volume V under consideration.

Expansion of the divergence operator in (2.4) leads to the equation

$$(\partial D/\partial C)(\nabla C)^2+D\nabla^2 C = \partial C/\partial t, \tag{2.4a}$$

which, in turn, reduces to an equation of simpler form by the Kirchhoff transformation [13, p. 11]

$$B = \int_0^C f(C')dC'. \tag{2.5}$$

When C satisfies Equation (2.4) then B satisfies the equation

$$D_o f(C)\nabla^2 B = \partial B/\partial t. \tag{2.6}$$

Specializing (2.5) and (2.6) to the linear case by setting $D = D_o$ and $f = 1$ yields the familiar homogeneous diffusion equation for constant diffusivity

$$D_o\nabla^2 C = \partial C/\partial t. \tag{2.6a}$$

The one-dimensional and linear predeposition diffusion problem is governed by the equation

$$D_o\partial^2 C/\partial z^2 = \partial C/\partial t \tag{2.7}$$

and by the boundary conditions and initial conditions

$$C(0,t) = C_s,\ t > 0, \tag{2.8a}$$

$$C(\infty,t) = 0,\ F(\infty,t) = 0, \tag{2.8b}$$

$$C(z,0) = 0. \tag{2.8c}$$

It has the well-known solution

$$C(z,t) = C_s\mathrm{erfc}(z/\sqrt{4D_o t}). \tag{2.9}$$

The root in the argument of the complementary error function defines the diffusion length; it is the rigorous measure of depth of the diffusion and will appear as a parameter in all diffusion profiles.

The formulation of the boundary conditions (2.8) and solution (2.9) tacitly assumes that the infinite half space $z > 0$ is a valid approximation for a semiconductor wafer which in reality is only a

fraction of a millimeter thick. It will become self evident in the course of the developments that such an approximation is indeed reasonable for all diffusion processes of interest.

The profile (2.9) represents the initial condition for the drive-in diffusion process which, for the time being, will be considered under the simplest boundary conditions

$$F(0,t) = -D\partial C(0,t)/\partial z = 0, \tag{2.10a}$$

$$C(\infty,t) = 0, \tag{2.10b}$$

$$C(z,0) = C_s \operatorname{erfc}(z/\sqrt{4D_p t_p}); \tag{2.10c}$$

it is assumed that there is no impurity flux through the xy-plane. The abridged notation $\partial C(0,t)/\partial z$ stands for the limit of the directed derivative as z approaches 0. The index p distinguishes parameters of the predeposition diffusion from those of the drive-in process which will either carry the index d or no index at all. The conditions (2.10) do not lead to a closed form solution. Traditionally the initial profile for the drive-in diffusion has been approximated by a layer of impurities at the surface containing the same number per unit area as the actual profile, expressed by

$$C(z,0) = C_s\sqrt{4D_p t_p/\pi}\,\delta(z). \tag{2.11}$$

This approximation does yield an analytic solution for the drive-in profile, given by

$$C(z,t) = C_s(2/\pi)\sqrt{D_p t_p/(Dt)}\exp[-z^2/(4Dt)]. \tag{2.12}$$

It is known as the delta function approximation and is good when the drive-in diffusion is much deeper than the predeposition diffusion. In many modern device structures this is not the case, however, and investigators have generally resorted to numerical solutions. Kennedy [14] has derived the two-step diffusion profile in a form convenient for numerical integration;

$$\left.\begin{aligned} C(z,D_p t_p,Dt) &= C_s(2/\sqrt{\pi})\int_{\zeta}^{\infty}\exp(-x^2)\,\mathrm{erf}(Kx)\,dx, \\ \zeta &= z/\sqrt{4D_p t_p+4Dt}, \qquad K = \sqrt{D_p t_p/(Dt)}. \end{aligned}\right\} \tag{2.13}$$

He computed profiles over three decades of the concentration for values of the parameter K ranging from 0.1 to 10. The final surface concentration, obtained from the integral between the limits zero and infinity, is expressible in closed form by

$$C(0,D_p t_p,Dt) = C_s(2/\pi)\arctan(K). \tag{2.14}$$

For Kx larger than 2 the error function can be approximated by unity. The integral becomes solvable in closed form and leads to a complementary error function for the tail of the profile. For values of Kx less than 0.2 the error function is almost proportional to its argument, and the profile approaches the Gaussian form. Much of the range of interest is not covered by these asymptotic solutions, and other approximations of simple form and extended range of applicability are required. To this end one needs an adequate set of functions which satisfy the diffusion equation and some of the boundary and initial conditions, in particular the boundary condition (2.8c). One might

first think of separation of variables techniques to find a set of solutions, but an inspection of the solutions (2.9) and (2.12) suggests other than product solutions. The causal character of the diffusion process points toward Laplace transform techniques. The solutions of Equation (2.7) in the transform domain are, by inspection,

$$C(z,s) = A(s)\exp(\pm z\sqrt{s/D}), \tag{2.15}$$

of which only those with the negative exponent satisfy the boundary conditions (2.8b). Table 2.1 lists some of the known inverse transforms, taken from the literature [15, formulae 29.3.86, 87, 89]. The index n runs from 0 over the positive integers. The symbol i^n stands for the nth integral of the complementary error function, and it is implied that $i^0\mathrm{erfc}(x) = \mathrm{erfc}(x)$ and that $i^{-1}\mathrm{erfc}(x) = -(2/\sqrt{\pi})\exp -x^2$. The functions H_n are the Hermite polynomials. It is worth noting that all functions (2.16) with odd n give zero concentration at the surface z = 0 while those with even n lead to vanishing flux at the surface and that all of them decay with the time. Furthermore, since differentiation changes the index by plus one, integration consequently will change the index by minus one. The total number of impurities per unit area can thus at once be determined;

$$\begin{aligned} Q_{0n} &= \int_0^\infty \frac{\exp -(z/\sqrt{4Dt})^2 H_n(z/\sqrt{4Dt})}{2^n \pi^{\frac{1}{2}} t^{\frac{1}{2}(n+1)}}\, dz \\ &= \begin{cases} 0, & n \text{ even} \\ D^{-\frac{1}{2}} H_{n-1}(0)/(2^{n-1}\pi^{\frac{1}{2}} t^{\frac{1}{2}n}), & n \text{ odd} \end{cases} \end{aligned} \tag{2.19}$$

Table 2.1. Inverse Laplace Transforms of $A(s)\exp-(2\sqrt{s/D})$

$A(s)$	$C(z,t)$	$-D\partial C(z,t)/\partial z$	
$s^{\frac{1}{2}n-\frac{1}{2}}$	$\dfrac{\exp-(z/\sqrt{4Dt})^2 H_n(z/\sqrt{4Dt})}{2^n \pi^{\frac{1}{2}} t^{\frac{1}{2}+\frac{1}{2}n}}$	$\dfrac{\sqrt{D}\exp-(z/\sqrt{4Dt})^2 H_{n+1}(z/\sqrt{4Dt})}{2^{n+1} \pi^{\frac{1}{2}} t^{1+\frac{1}{2}n}}$	(2.16.n)
$s^{-\frac{1}{2}n-1}$	$(4t)^{\frac{1}{2}n} i^n \operatorname{erfc}(z/\sqrt{4Dt})$	$\sqrt{D}(4t)^{\frac{1}{2}n-\frac{1}{2}} i^{n-1}\operatorname{erfc}(z/\sqrt{4Dt})$	(2.17.n)
$\dfrac{1}{s+a\sqrt{s}}$	$\exp(az/\sqrt{D}+a^2 t)\cdot$ $\cdot\operatorname{erfc}(a\sqrt{t}+z/\sqrt{4Dt})$	$\sqrt{D}\exp(az/\sqrt{D}+a^2 t)\cdot$ $\cdot[(1/\sqrt{\pi t})\exp-(a\sqrt{t}+z/\sqrt{4Dt})^2 -$ $-a\cdot\operatorname{erfc}(a\sqrt{t}+z/\sqrt{4Dt})]$	(2.18)

The functions (2.17) all start with $C(z,0) = 0$, and the concentration increases with time. When $n = 0$ the flux at the surface decreases with time; it is constant when $n = 1$, and it increases with time when $n > 1$. Since neither of the two sets of functions is orthogonal, simple expansion algorithms cannot be applied. However, since the goal here is to find approximations of simple form, only a few functions will be used to describe a profile; and the lack of orthogonality is not a severe handicap.

Returning now to the problem of approximating the profile of the simple drive-in diffusion, defined by the conditions (2.10), one can either approximate the initial condition (2.10c) with functions which precisely satisfy the two boundary conditions, or one can

approximate the boundary condition (2.10a) with functions satisfying (2.10b) and (2.10c) exactly. The delta-function approximation is an example of the first approach. It uses the total number of impurities in the profile as the single approximation criterion[1] and approximates the initial profile with the solution (2.16.0) for $t = t_d$. Using the same function, one can immediately think of a better approximation with two parameters, introducing a time shift in addition and using the surface concentration as an additional approximation criterion,[2]

$$\tilde{C}(z,t) = C_d\sqrt{t_o/(t_o+t)}\exp-[z/\sqrt{4D(t_o+t)}]^2, \tag{2.20}$$

$$\tilde{C}(0,0) = C(0,t_p), \quad \tilde{Q}(0,0) = Q(0,t_p). \tag{2.21}$$

The condition on the surface concentration leads to $C_d = C_s$, by comparison of (2.10c), (2.20), and (2.21). The condition on Q equates the integrals

$$\int_0^\infty \operatorname{erfc}(z/\sqrt{4D_p t_p})dz = \int_0^\infty \exp-(z/\sqrt{4Dt_o})^2 dz \tag{2.22}$$

and leads to

$$t_o = (2/\pi)^2 D_p t_p/D. \tag{2.23}$$

The quality of this approximation is illustrated in Figure 2.1, where errors of two different approximations and exact profiles are plotted for a number of drive-in parameters. The approximate profile, (2.20),

---

1. Approximation criterion: A quantity which is required to have the same value in the approximation and in the exact profile.

2. The tilde henceforth indicates approximations.

deviates no more than 10% (3%) from the exact one over six decades of concentration, if $D_d t_d \geq 30\ D_p t_p$ ($100\ D_p t_p$). The addition of higher order functions of the type (2.16.n), preferably those with even n which conserve Q, would improve this approximation. Instead of doing this, the alternate approach--approximating the boundary condition, (2.10a), with functions which satisfy (2.10c)--shall now be demonstrated.

The predeposition diffusion profile evolves in time like the function (2.17.0), and the flux profile is described by

$$F_p(z,t_p) = C_s\sqrt{D_p/\pi t_p}\exp-(z/\sqrt{4D_p t_p})^2. \tag{2.24}$$

This function can be transformed to the time scale of the drive-in diffusion, which has its origin at a time point equivalent to t_p. The transformed relations are, after some thought,

$$C(z,t_d) = C_s\mathrm{erfc}[z/\sqrt{4D_d(t_1+t_d)}], \tag{2.25}$$

$$F(0,t_d) = C_s\sqrt{D_d/[\pi(t_1+t_d)]}, \tag{2.26}$$

$$t_1 = D_p t_p/D_d, \tag{2.27}$$

During the drive-in interval, $(0,t_2)$, the drive-in regime replaces the predeposition regime, i.e., the flux at the surface z = 0 is forced to zero. The function (2.25) must now be augmented by approximation functions which cancel the flux (2.26) during the drive-in period. Since Equation (2.25) describes the initial concentration, (2.10c), exactly, the added functions should all vanish at $t_d = 0$. All functions of the set (2.17) exhibit this property.

The approximation needs only to be good in the interval, $(0,t_2)$, because the diffusion process stops at t_2.

Let the first approximation be in terms of the function (2.17.1). Since the proper boundary condition at the surface $z = 0$ cannot be met point by point in this way, the next best fit is "on the average," i.e.,

$$\int_0^{t_2}[F(0,t_d)+\tilde{F}(0,t_d)]dt_d = C_s\sqrt{D_d/\pi}\int_0^{t_2}(t_1+t_d)^{-\frac{1}{2}}dt_d+a\sqrt{D_d}\int_0^{t_2}dt_d = 0. \tag{2.28}$$

The constant

$$a = -2C_s(\sqrt{t_1+t_2}-\sqrt{t_1})/\sqrt{\pi t_1} \tag{2.29}$$

satisfies this condition, and the first approximation of the drive-in profile at $t_d = t_2$ becomes

$$\tilde{C}(z,t_2) = C_s\{\operatorname{erfc}[z/\sqrt{4D(t_1+t_2)}]-4[(\sqrt{t_1+t_2}-\sqrt{t_1})/\sqrt{\pi t_2}]\cdot$$

$$\cdot i^1\operatorname{erfc}(z/\sqrt{4Dt_2})\}. \tag{2.30}$$

The addition of the function (2.17.2) would provide for one additional free parameter. Consider, however, at once a further expansion which also includes the function (2.17.3), with the free parameters determined such that the boundary condition at the surface is exactly met at both end points of the drive-in time interval and on the average. Dropping the index, d, for simplicity, one finds that the flux function assumes the form

$$F(0,t) = C_s\sqrt{D/\pi}(1/\sqrt{t_1+t}+a_1+a_2\sqrt{t}+a_3 t) \tag{2.31}$$

and the three conditions

$$\tilde{F}(0,0) = 0; \quad \tilde{F}(0,t_2) = 0; \quad \int_0^{t_2} \tilde{F}(0,t)dt = 0 \tag{2.32}$$

determine the three constants, a_i. Straightforward algebra yields

$$\left.\begin{aligned} a_1 &= -1/\sqrt{t_1}, \\ a_2 &= (3/t_2)[4\sqrt{t_1/t_2}+\sqrt{t_2/t_1}+\sqrt{t_2/(t_1+t_2)}-4\sqrt{(t_1+t_2)/t_2}], \\ a_3 &= (2/t_2)(6\sqrt{t_1+t_2}/t_2-6\sqrt{t_1}/t_2-2/\sqrt{t_1+t_2}-1/\sqrt{t_1}). \end{aligned}\right\} \tag{2.33}$$

The approximation of the profile at the end of the drive-in diffusion becomes

$$\tilde{C}(z,t_2) = C_s\{\mathrm{erfc}[z/\sqrt{4D(t_1+t_2)}]+2a_1\sqrt{t_2/\pi}\,i^1\mathrm{erfc}(z/\sqrt{4Dt_2})+$$

$$+2a_2 t_2 i^2\mathrm{erfc}(z/\sqrt{4Dt_2})+8a_3 t_2\sqrt{t_2/\pi}\,i^3\mathrm{erfc}(z/\sqrt{4Dt_2}). \tag{2.34}$$

As long as the diffusion length of the drive-in cycle is not more than 30 (10) times larger than that of the predeposition cycle, this approximation deviates by no more than 12% (6%) from the exact profile, over a range of six decades in concentration.

This and the preceding approximation, then, cover all two-step diffusion profiles with an accuracy of 10% or better over the range of practical significance in the concentration. The delta-function approximation, on the other hand, is up to 50% in error, even if the drive-in diffusion length is a hundred times larger than the

predeposition diffusion length. Figure 2.1 illustrates a number of profile curves of two-step diffusions, over six decades, for values of K between 0,1 and 10. It also shows some curves of relative errors of approximations. The broken and solid lines indicate errors of Equations (2.20) and (2.34) respectively.

The two approximation examples suggest that the selection of appropriate approximation criteria is more a matter of intuition than a rigorous method. The ideas for higher order approximations would soon be exhausted and the determination of the parameters would become exceedingly tedious. However, the improvement in accuracy to be expected from the addition of further terms would hardly justify the effort.

In closing this section it might be remarked that the functions of Table 2.1 allow for considerably more flexibility in approximating diffusion profiles in closed form than the literature on solid state devices might suggest. One very important class of profiles, those resulting from drive-in diffusion with reoxidation, however, is not amenable to this type of approximation. Huang and Welliver [16] have attempted an approximate closed form solution based on the delta function approximation. Although their profiles show the expected overall features, the solutions suffer from vast nonconservation of the impurities at the moving silicon dioxide-silicon interface. The same problem arises with comparable severity in approximations using up to four of the functions from Table 2.1. These functions satisfy boundary conditions at the initial silicon surface, but are ill-suited for describing conditions of a moving boundary.

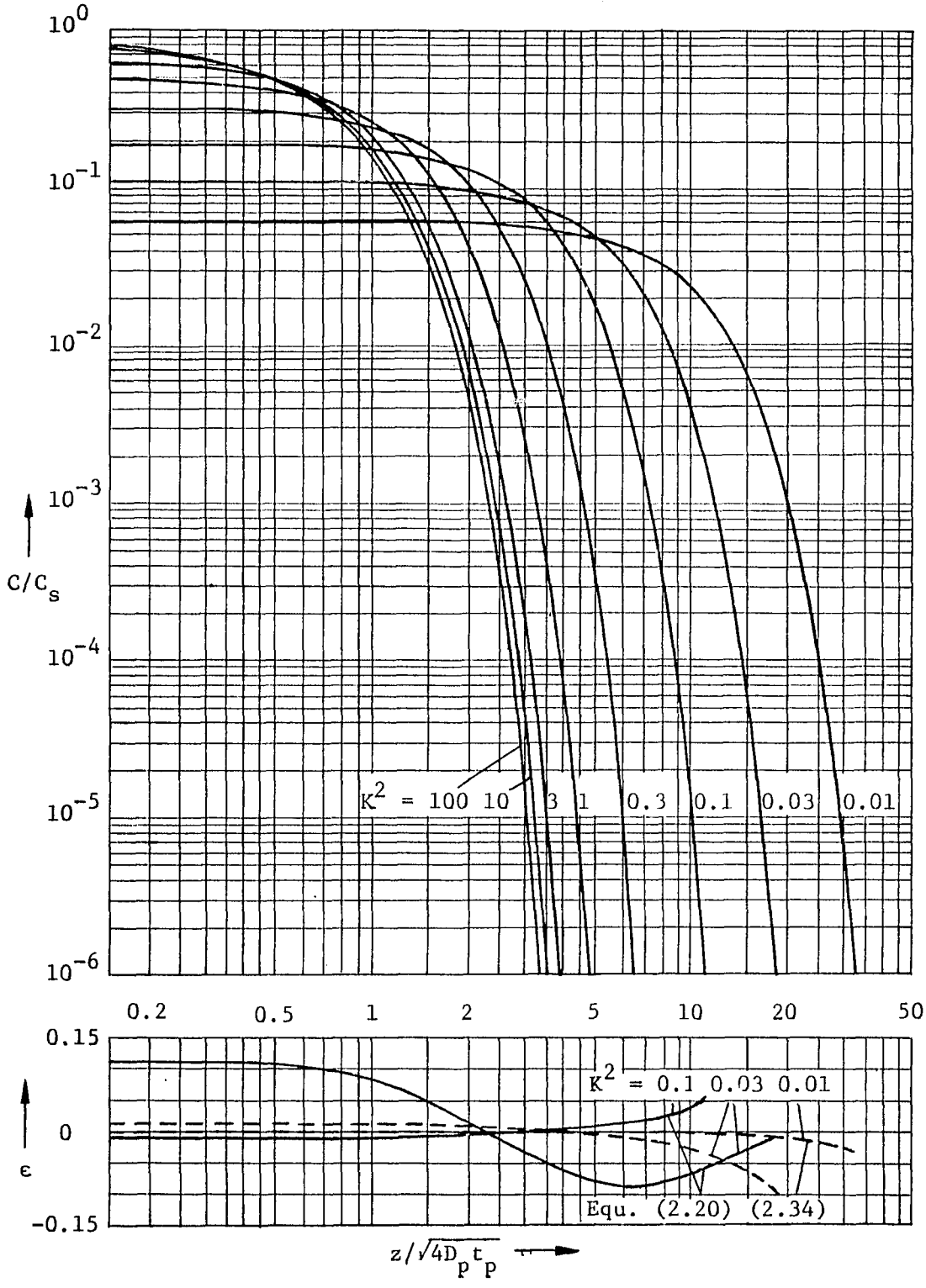

Figure 2.1. Accurate Profile of One-Dimensional Two-Step Diffusion and Relative Errors, ϵ, of Two Approximations

2.4 Three-Dimensional Impurity Profiles from Rectangular Diffusion Windows--Methods for Numerical Calculation

As the diffusion windows become small, it is necessary to consider three-dimensional profiles or, for long and slim geometries, at least two-dimensional profiles.

The three-dimensional boundary value problem of the predeposition diffusion is specified by the diffusion equation (2.6a) and the boundary conditions

$$C(\vec{r}_w,t) = C_s, \quad t > 0, \tag{2.35a}$$

$$\partial C(\vec{r}_s,t)/\partial\hat{n} = 0, \tag{2.35b}$$

$$\nabla C(\vec{r}_\infty,t) = 0, \tag{2.35c}$$

$$C(\vec{r}_\infty,t) = 0. \tag{2.35d}$$

C_s still denotes the solid solubility; $\vec{r}_w$ stands for the position vectors of points in the xy-plane inside the window; $\vec{r}_s$ indicates points in the xy-plane outside of the window; and $\vec{r}_\infty$ locates points which are exceedingly far away from the window. The abridged notation $\partial C(\vec{r}_s,t)/\partial\hat{n}$ stands for the limit, as $\vec{r}$ approaches $\vec{r}_s$, of the directed derivative of C along the outward normal to the surface, and similarly $C(\vec{r}_\infty,t)$ and $\nabla C(\vec{r}_\infty,t)$ respectively indicate the limits of the concentration and its gradient as $\vec{r}$ approaches infinity. The caret indicates a unit vector.

The initial condition for the predeposition is

$$C(\vec{r},0) = 0, \tag{2.36}$$

The profile created by the predeposition diffusion,

$$C(\vec{r},t_p) \equiv C_p(\vec{r}) = C(\vec{r},t_d=0), \tag{2.37}$$

in turn, represents the initial condition for the drive-in diffusion. The boundary conditions for the simple drive-in diffusion are those of Equations (2.35) with the exception that (2.35b) replaces (2.35a) for the points inside the window.

None of these problems have closed-form solutions, and one has to resort to numerical iteration techniques or to approximations for the calculation of the profiles.

The finite difference techniques are probably the foremost iterative methods because of their apparent universality and the ease with which they can be implemented in computer programs. Their value for the calculation of diffusion profiles can be judged by the accuracy achievable over the large range of concentration required and the computational expense--number of grid points and number of iteration steps--necessary. Once again, the example of the one-dimensional predeposition diffusion can highlight the behavior and problems of the method.

Consider a space-time grid with points, i,j, located at the coordinates $x_i,t_j = i\Delta x, j\Delta t$, and let the values of the exact and approximate solutions be denoted by $u_{i,j} = u(x_i,t_j)$ and $v_{i,j} = v(x_i,t_j)$ respectively. Let the initial and boundary conditions be $u_{i,0} = 0$ $(i \neq 0)$ and $u_{0,j} = 1$, and consider the case of unity diffusivity. This leads to the diffusion equation

$$u_{xx} = u_t, \tag{2.38}$$

subscripts denoting derivatives.

Two kinds of finite difference schemes are distinguished. In the explicit method the value $v_{i,j+1}$ is calculated from the three values $v_{i-1,j}$, $v_{i,j}$, and $v_{i+1,j}$, while the implicit methods provide equations linking the six values $v_{i-1,j}$, $v_{i,j}$, $v_{i+1,j}$, $v_{i-1,j+1}$, $v_{i,j+1}$, and $v_{i+1,j+1}$. The explicit scheme allows one to determine the solution at the time $t + \Delta t$ from that at the time t point by point with an algorithm, while the implicit schemes require the simultaneous solution of a system of equations in each step. The stability criteria of the methods are outlined in books on numerical methods, for example by Carnahan, Luther, and Wilkes [17, pp. 431-433] and are retraced here to highlight the problem of estimating the accuracy of the result. The algorithm of the explicit method is

$$v_{i,j+1} = v_{i,j} + p(v_{i-1,j} + v_{i+1,j} - 2v_{i,j}), \tag{2.39}$$

$$p = \Delta t/(\Delta x)^2. \tag{2.40}$$

With p larger than $\frac{1}{2}$ Equation (2.39) will not model a diffusion process because a disturbance of magnitude +1 at the point i,j will produce disturbances equal to 1-2p at the point i,j+1 and equal to p at the points $i\pm1$,j+1, etc. The sum of the magnitudes of these disturbances grows with each step, leading to an ever increasing oscillatory instability in the computation.

A Taylor expansion about the point i,j eventually leads to an estimate of the error. It relates the values of the exact solution u

on adjacent grid points:

$$u_{i,j+1} = u_{i,j}+(\Delta t)u_t+(\Delta t)^2u_t/2!+(\Delta t)^3u_{ttt}/3! \ldots, \tag{2.41}$$

$$u_{i\pm1,j} = u_{i,j}\pm(\Delta x)u_x+(\Delta x)^2u_{xx}/2!+(\Delta x)^3u_{xxx}/3! \ldots. \tag{2.42}$$

Solving the equation

$$u_{i+1,j}-u_{i-1,j} = 2[u_{i,j}+(\Delta x)^2u_{xx}/2!]+\ldots \tag{2.43}$$

for u_{xx} and replacing u_t in (2.41) with it leads to

$$u_{i,j+1} = u_{i,j}+p(u_{i+1,j}+u_{i-1,j}-2u_{i,j})+\Delta e_{i,j}, \tag{2.44}$$

$$\Delta e_{i,j} = (\Delta t)^2u_{tt}/2!+(\Delta t)^3u_{ttt}/3!-$$

$$-2p[(\Delta x)^4u_{4x}/4!+(\Delta x)^6u_{6x}/6!] \ldots. \tag{2.45}$$

Defining the errors as the quantities $e_{i,j} = v_{i,j}-u_{i,j}$ leads to the difference between (2.39) and (2.44) in the form

$$e_{i,j+1} = e_{i,j}+p(e_{i+1,j}+e_{i-1,j}-2e_{i,j})-\Delta e_{i,j}. \tag{2.46}$$

Provided that the higher order derivatives of u exist, the relationships

$$u_{tt} = (u_t)_t = u_{xxt} = (u_t)_{xx} = u_{xxxx}, \tag{2.47}$$

and similarly $u_{ttt} = u_{6x}$ are true. Therefore the first and third term of the right hand side of (2.45) can be combined;

$$(\Delta t)^2u_{tt}/2!-2p(\Delta x)^4u_{4x}/4! = \tfrac{1}{2}(\Delta t)^2[1-1/(6p)]u_{tt}, \tag{2.48}$$

and they cancel when p = 1/6. In that case the error increment reduces to

$$\Delta e_{i,j} = (\Delta t)^3 u_{ttt}/3! - (\Delta x)^6 u_{6x}/(3 \cdot 6!) \ \ldots \tag{2.49}$$

which is in lowest order proportional to $(\Delta t)^3$ rather than to $(\Delta t)^2$. In view of (2.47) the error increment can then be approximated by

$$\Delta e_{i,j} = (1/15)(\Delta t)^3 u_{ttt}. \tag{2.50}$$

Now, in order to evaluate the error $e_{i,j}$, the whole pyramid of previous errors, down to the first step of the iteration, would have to be evaluated. To this end it would be necessary to know the derivative u_{ttt} which, in practice, cannot be known any more than the solution itself. Thus it is quite impossible to obtain a generally valid estimate of the error. On the other hand, impurity diffusion in integrated devices is always characterized by the process starting with well localized sources and by steps in the initial condition. The solutions will be generic to complementary error functions, the three- and two-dimensional solutions decaying more rapidly with distance than the one-dimensional ones. Returning now to the profile of the one-dimensional predeposition problem with the exact solution (for D=1),

$$u = \mathrm{erfc}(x/\sqrt{4t}) \tag{2.51}$$

one can interpret the time variable as a parameter which scales the error function with respect to the x-variable. Considering a normalized profile erfc(ξ) where

$$\xi = x/\sqrt{4t}, \tag{2.52}$$

the range of ξ between 0 and 3.5 covers a range of six orders of magnitude of the concentration, and it is desired to approximate the error function to good accuracy over this range. The initial and boundary conditions are

$$v_{i,0} = 0; \quad i = 1,2,\ldots; \quad v_{0,j} = 1; \quad j = 0,1,2,\ldots, \tag{2,53}$$

where, strictly speaking, the index i is unbounded. This problem, though, will remove itself in due course. In view of Equations (2,52) and (2,40) the mesh ratio p = 1/6 determines the arguments of the points i,j as

$$\xi_{i,j} = i\Delta x/\sqrt{4j\Delta t} = i\sqrt{3/(2j)} \tag{2.54}$$

In the first iteration step one therefore tries to calculate the error function at points $\xi_i = i\sqrt{3/2}$. At the second iteration the points are spaced by $\sqrt{3/4}$, etc. On the other hand, the derivation leading to the error increment, $\Delta e_{i,j}$, tacitly assumed that the grid points would be closely spaced with respect to the features of the solution curve such that the piecewise linear approximation would represent a good fit to the curve at any time step. Because of the step in the initial condition, as expressed by Equations (2.53), this assumption cannot hold for the early steps of the iteration process. Hence no reliable error estimate can be obtained with the aid of Equation (2,50), and an inspection of the results of the iteration is the only avenue left for obtaining information about the convergence of the method. Table 2,2 summarizes the results of the first four iterations.

Table 2.2. Values $v_{i,j}$

j	i = 0	1	2	3	4	5
0	v = 1	0	0	0	0	0
1	1	0.167	0	0	0	0
2	1	0.278	0.0278	0	0	0
3	1	0.357	0.0648	0.00463	0	0
4	1	0.415	0.1034	0.0139	0.00077	0

The relative errors, defined by

$$\varepsilon_{i,j} = (v_{i,j}-u_{i,j})/u_{i,j}, \tag{2.55}$$

for the first 50 iterations in the range $0 \leq \xi \leq 3.5$ are displayed in Figure 2.2. Every fifth step is represented by a black dot, while circles indicate steps in between. The errors are only shown for the grid points with $i \leq j$. For all other points the relative error equals -1 because $v_{i,j}$ is zero as indicated in Table 2.2. All errors converge toward zero; after 25 steps the maximum error in the range indicated is 0.15, and after 50 steps it is 0.08. At the twenty-fifth step the grid has four points per diffusion length.

To calculate the profile from a rectangular diffusion window of size 8 μm by 4 μm at a diffusion length of 0.2 μm, a grid spacing of 0.05 μm would have to be used in order to achieve an accuracy of about ten to twenty per cent in 25 steps. The modeled region should be at least five diffusion lengths deep and should exceed the window by the

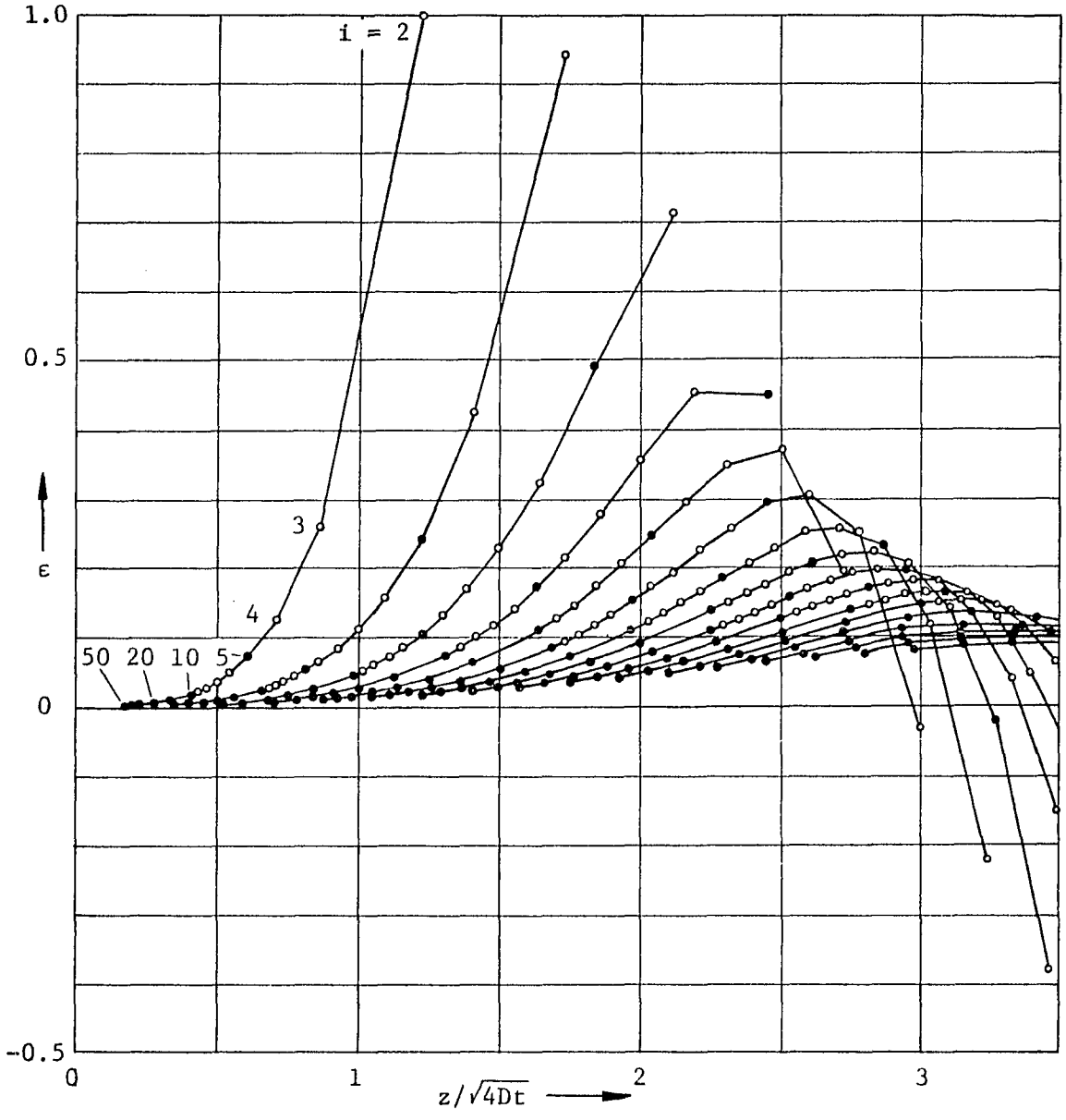

Figure 2.2. Relative Error, ε, at the Nodal Points of a Finite Difference Calculation of the One-Dimensional Predeposition Profile at the ith Step

same amount in all directions, resulting in a block of size 10 μm by 6 μm by 1 μm. Due to symmetry it suffices to calculate the profile in one quarter of this region, but the lattice would still contain 120,000 points. Halving the mesh size would lead to a million points.

This exercise demonstrates that the finite difference schemes require considerable computational expense for calculations of only moderate accuracy and that it is worth the effort to look for more efficient ways to calculate impurity profiles.

The next method in line is the finite element method. It is based on the variational principle and is said to require fewer grid points than the finite difference scheme because it works also with variable grid size.

The calculus of variation has led to new insights in many instances and is a powerful and general formulation for describing and relating many physical phenomena. On the other hand it is conceptually intricate enough to lead an unsuspecting researcher astray if inappropriately applied.

Morse and Feshbach [9, p. 313] discuss the variational concepts in quite general terms and formulate the more important field problems in a canonical form derived from Hamiltonian mechanics. They, and others, claim that a variational formulation of the diffusion equation can be found in terms of the diffusing variable and its adjoint as well as their derivatives in space and time. If, e.g., the variable C satisfies Equation (2.4) then the adjoint variable $\bar{C}$ solves the adjoint equation

$$\nabla\cdot(D\nabla\bar{C}) = -\partial\bar{C}/\partial t. \qquad (2.56)$$

The Hamiltonian formulation can then be applied to derive the Lagrangian density

$$L = -D\nabla C\cdot\nabla\bar{C}-\tfrac{1}{2}(\bar{C}\partial C/\partial t-C\partial\bar{C}/\partial t), \tag{2.57}$$

the state variables C and $\bar{C}$ being functions of the coordinates x, y, z, and t. The functional, Ω, is the integral over the volume of the diffusion problem and over the time interval of the diffusion process,

$$\Omega = \int_V \int_T L d\vec{r}dt. \tag{2.58}$$

The true solution $C(\vec{r},t)$ extremizes this functional and thus its variation must vanish. This condition is expressed by the Euler-Lagrange equation which, in this case, is of the form

$$\begin{aligned} 0 &= \{\partial/\partial\bar{C}-\nabla\cdot[\partial/\partial(\nabla\bar{C})]-(d/dt)[\partial/\partial(\partial\bar{C}/\partial t)]\}L \\ &= -\tfrac{1}{2}\partial C/\partial t+\nabla\cdot(D\nabla C)-\tfrac{1}{2}\partial C/\partial t = \nabla\cdot(D\nabla C)-\partial C/\partial t, \end{aligned} \tag{2.59}$$

for the variation with respect to $\bar{C}$. Indeed, that concentration profile which makes the functional stationary also satisfies the diffusion equation. Similarly the variation of Ω with respect to C recreates the adjoint diffusion equation. The functional is only stationary when both variations vanish simultaneously, i.e., when C satisfies Equation (2.4) and at the same time $\bar{C}$ satisfies Equation (2.57). So far nothing has transpired about the nature of the adjoint variable, and it has not been identified by a physical variable of the system. To shed some light into the situation, consider heat diffusion along a piece of wire which is heated or cooled at its ends.

The variational formulation would solve a problem of the type "given an initial and a final temperature distribution, how must one

heat or cool the ends of the wire with time in order to evolve from the initial to the final temperature distribution." The nature of the diffusion process is such that, as time passes, it tends to smooth out all variations of the temperature in a source free region. At a local maximum of the temperature, its divergence is positive; and to satisfy the diffusion equation, its time derivative must be negative, thus reducing the maximum. By the same line of argument the local minima will become shallower with time. With this in mind it is clear that there is, for instance, no way to achieve a final condition of the temperature distribution of the wire of length L given as

$$T_f = \begin{cases} T_o, & L/3 < x < 2L/3 \\ \\ 0, & 0 < x < L/3, \quad 2L/3 < x < L \end{cases} \tag{2.60}$$

by heating or cooling the wire from its ends. Yet, the variational formulation outlined so far does not formally exclude such final conditions and hence must be able to accommodate them. This is achieved by introducing the adjoint variable. The nature of the adjoint diffusion process is opposite to the real process; it progressively peaks the adjoint temperature up at local maxima and down at local minima as the time evolves. By introducing the adjoint process into the system, one loses the distinction between physically realizable and nonrealizable solutions. Furthermore, the diffusion problems are posed in a different way than required by the variational formulation. One wishes to find the final profile, given the initial profile and the boundary conditions. Gurtin [18] has shown how a

variational formulation for such problems can be derived for the linear diffusion equation and for the wave equation. To this end he considers the problem defined by Equation (2.6a) with a general initial condition

$$C(\vec{r},0) = C_o(\vec{r}) \tag{2.61}$$

and boundary conditions typically like (2,35). He then goes on to prove that such a problem leads to a functional of the form

$$\Omega = \int_V (C*C+D_o\nabla C*\nabla C+2C_o*C)d\vec{r} - 2\int_W (D_o*C_s*C)d\sigma, \tag{2.62}$$

the variation of which will vanish precisely if C is the solution of the diffusion problem. The integrals are respectively over the diffusion region and over the source regions, i.e., over the window surface. The stars stand for convolutions, the convolutions of the gradients being defined as

$$\nabla C*\nabla C = \int_0^t \nabla C(\tau)\cdot\nabla C(t-\tau)d\tau. \tag{2.63}$$

In this formulation the adjoint variable has been eliminated by explicitly invoking the causality of the system via the convolution integrals.

To study the merits of the variational method for the calculation of impurity profiles, consider once more the one-dimensional predeposition process and its approximation by the finite element method.

The subdivision of the diffusion region into finite elements--line segments in one dimension, triangles or rectangles in two dimensions, and tetrahedra or parallelepipeds and the like in three dimensions--represents the first step of the procedure. The global

region has to be finite; in the one-dimensional example it is to be subdivided into N line elements; the ith element lying between the points i-1 and i. The solution is next approximated in each element by a simple function, preferably by a low order polynomial, such that the integrals of the functional can be calculated in closed form within each element, and such that at least the function is continuous across the boundary of the elements. The mechanics of constructing the finite elements and of approximating the functions are well described in the literature, e.g., by Zienkiewicz [11].

In the one-dimensional problem, for a linear approximation, the temporal interpolation functions in the interval $0,\Delta t$, are

$$r = (\Delta t - t)/(\Delta t), \quad s = t/(\Delta t); \tag{2.64}$$

the spatial interpolation functions for the ith element are

$$u_i = (x_i - x)/(x_i - x_{i-1}), \quad v_i = (x - x_{i-1})/(x_i - x_{i-1}); \tag{2.65}$$

and they approximate the function $C(x,t)$ by the function

$$C_i(x,t) = [C(x_{i-1},0)r(t)+C(x_{i-1},\Delta t)s(t)]u_i(x)$$

$$+[C(x_i,0)r(t)+C(x_i,\Delta t)s(t)]v_i(x). \tag{2.66}$$

The approximation is in terms of the values of the variable C at the lattice points of the space-time lattice.

The convolution in (2.62) operates only on the functions r and s, while the spatial integration only operates on u and v. The convolution integrals are

$$r*r = s*s = \tfrac{1}{2}r*s = \Delta t/6 \tag{2.67}$$

and the spatial integrals are

$$\int_{\Delta x} dx = 2\int_{\Delta x} u_i dx = 2\int_{\Delta x} v_i dx = 3\int_{\Delta x} u_i^2 dx = 3\int_{\Delta x} v_i^2 dx$$

$$= 6\int_{\Delta x} u_i v_i dx = \Delta x. \tag{2.68}$$

The contribution of the ith element to the functional consists of products of the integrals of the interpolation functions and two parameters from the set $C(x_i,0)$, $C(x_i,\Delta t)$, $C(x_{i-1},0)$, $C(x_{i-1},\Delta t)$. The complete functional will be a sum of such terms, each containing a product of exactly two of the parameters. The extremization of the functional with respect to $C(x,t)$ is performed by setting the partial derivatives with respect to each of the parameters equal to zero. These derivatives are linear combinations of the variables $C(x_i,0)$ and $C(x_i,\Delta t)$ and thus one obtains a system of simultaneous linear equations for these parameters. It can be written in the form

$$[A]\vec{C}(\Delta t) = [B]\vec{C}(0). \tag{2.69}$$

The vector

$$\vec{C}(\Delta t)^T = [C(x_1,\Delta t), C(x_2,\Delta t), \ldots\ C(x_{N-1},\Delta t)] \tag{2.70}$$

contains all unknown parameters, while the vector

$$\vec{C}(0)^T = [C(0,0), C(x_1,0), \ldots\ C(x_N,0), C(0,\Delta t), C(x_N,\Delta t)] \tag{2.71}$$

contains all the parameters known by virtue of the initial and boundary conditions. The solution entails the inversion of the matrix [A] once. Thereafter the vector of unknowns is calculated for the first time

step. These results are next used as the initial condition for the second time step, etc.

Figure 2.3 illustrates the performance of this iteration scheme. The relative errors are plotted for the first fifty steps in exactly the same way as in Figure 2.2. The accuracy of the finite element method is considerably inferior to that of the finite difference method; at low concentrations the errors after the fiftieth iteration are still as much as 90%, while the finite difference scheme showed less than 10% error in the same region.

The crucial problem with the iterative methods is the propagation of the errors from step to step which inevitably leads to the loss of all accuracy in the regions of low concentration. For this reason the iterative schemes are not suitable for calculating impurity profiles. Green's function techniques appear more promising. The Green's functions for the diffusion of impurities through a window are known in closed form and some of the integrals can be solved in closed form also. The remaining numerical evaluations, finally, do not suffer from the error propagation problems encountered with the iterative techniques.

The concept underlying the integral representation of the problem is simple enough: The concentration at a point $\vec{r}$ at the time t is the superposition of the contributions from all sources which have been active at any time t' before t at any point $\vec{r}'$ in the space under consideration. In particular, the contribution due to an impulsive source of strength unity is the Green's function $G(\vec{r},t|\vec{r}',t')$.

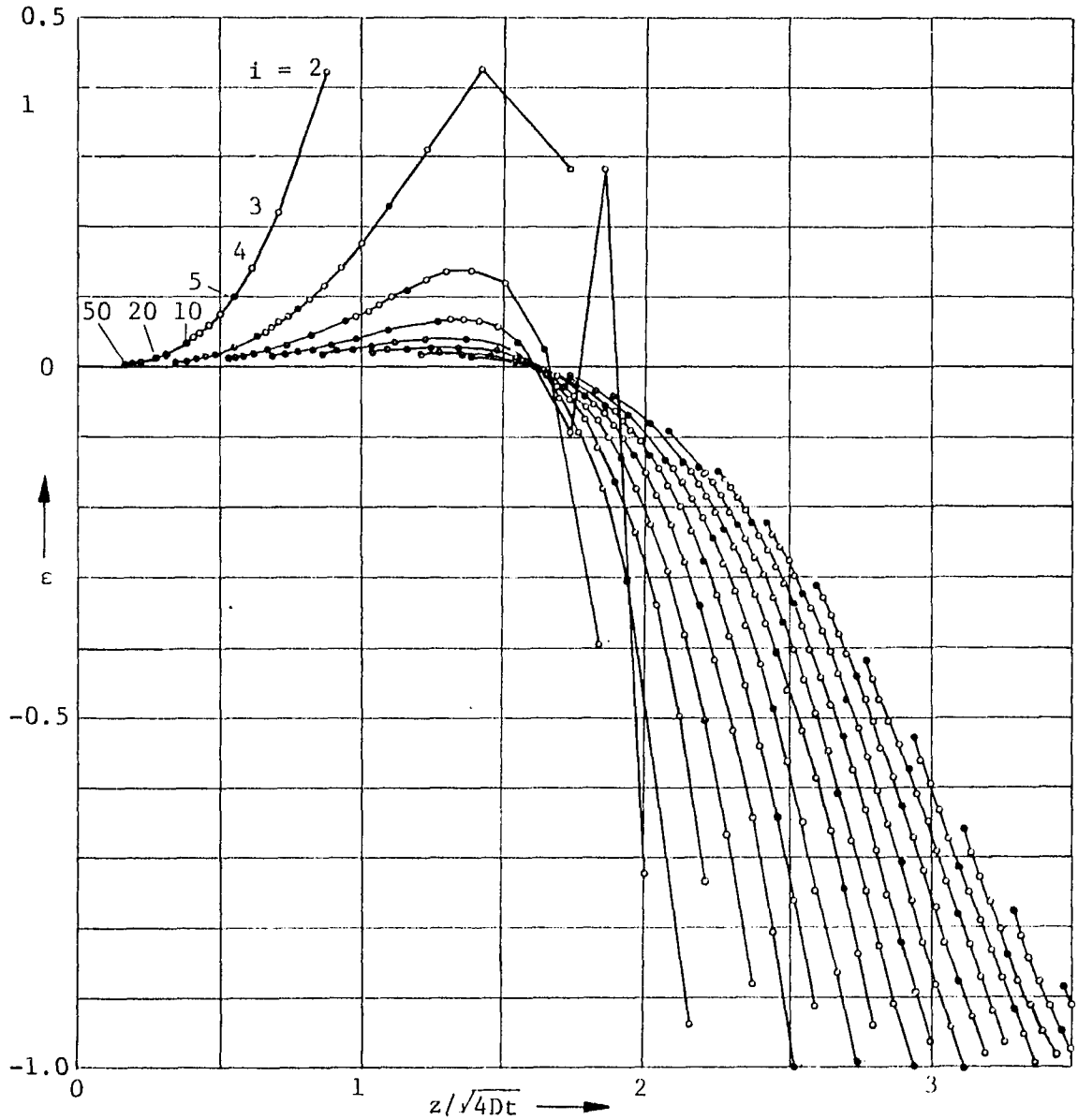

Figure 2.3. Relative Error, ε, at the Nodal Points of a Finite Element Calculation of the One-Dimensional Predeposition Profile at the ith Step

Stackgold [19, Chapter 7] gives a sufficiently rigorous and lucid derivation of the integral formulation based on the free space Green's function for the diffusion equation. The quintessence of the analysis is contained in the equations

$$C(\vec{r},t) = \int_0^t\int_{V'} G_o(\vec{r},t|\vec{r}',t')S(\vec{r}',t')d\vec{r}'dt'+$$

$$+\int_{V'} G_o(\vec{r},t|\vec{r}',0)\, C\,(\vec{r}',0)d\vec{r}'-$$

$$-D_o\int_{-\infty}^t\int_\sigma \{G_o(\vec{r},t|\vec{r}',t')\partial C(\vec{r}'_\sigma,t')/\partial\hat{n}'-$$

$$-[\partial G_o(\vec{r},t|\vec{r}'_\sigma,t')/\partial\hat{n}]C(\vec{r}'_\sigma,t')\}d\sigma'dt', \qquad (2.72)$$

$$G_o(\vec{r},t|\vec{r}',t') = [4\pi D_o(t-t')]^{-3/2}U(t-t')\cdot$$

$$\cdot\exp-\{(\vec{r}-\vec{r}')^2/[4D_o(t-t')]\}, \qquad (2.73)$$

where $S(\vec{r}',t')$ represents the density of sources (e.g., measured in $cm^{-6}s^{-1}$); V denotes the volume under consideration (in the problem at hand the half space $z \geq 0$); σ stands for the surface enclosing the volume V; and U is the unit step function, reflecting the causality of the diffusion process. All other notation is concurrent with previous definitions.

Some interpretation of Equation (2.72) with respect to the half space $z \geq 0$ is in order at this point. To this end consider a source density $S(\vec{r}',t')$ in an elemental volume $d\vec{r}'$ located at $\vec{r}' = (x',y',z'>0)$.

The first integral of (2.72) gives the contribution of these sources to the concentration. The second integral determines the contribution from the initial profile at the time $t' = 0$. Because of the free space Green's function, (2.73), these two contributions generate a continuous profile over all space without regard to the limitation of the volume to $z \geq 0$. The last integral accounts for the boundary condition, which requires the vertical component of the flux through the xy-plane to be zero at all times.

The first term of the surface integral compensates for the extraneous flux created by the volume integrals. This flux is given as

$$\vec{F}_{-z} = -D_o \partial C(\vec{r}_\sigma, t)/\partial z = D_o \partial C(\vec{r}_\sigma, t)/\partial \hat{n}, \tag{2.74}$$

and it is compensated for by a surface source density $-\vec{F}_{-z}$ contained in the first term of the surface integral. The resulting profile has a step of height $C(\vec{r}_\sigma, t)$ in the xy-plane, because

$$C(z=0_+, t) = C(\vec{r}_\sigma, t); \quad C(z=0_-, t) = 0, \tag{2.75}$$

This step has to be supported by a layer of doublet sources of strength $D_o C(\vec{r}_\sigma, t)$, polarized in the $+z$ direction. The contributions of these sources are given by the second term in the surface integral.

The simple geometry of the upper half space makes possible a considerably simpler formulation in terms of images. If each source in the free space is supplemented by its mirror image with respect to the xy-plane, and if further the initial concentration is chosen symmetrical with respect to the xy-plane, then the flux through the xy-plane vanishes for all time. (Since the physical system consists of the half

space $z \geq 0$ only, one is free to choose the sources and initial conditions in this manner.) The surface integral in Equation (2.72) will be zero for such a system, and only the first two integrals will remain.

Instead of introducing the images of the sources and initial condition, one can augment the Green's function by its mirror image, thus obtaining the half space Green's function

$$G(\vec{r},t|\vec{r}',t') = [4\pi D(t-t')^{-3/2}U(t-t')\exp[\frac{-(x-x')^2-(y-y')^2}{4D(t-t')}]$$

$$\cdot\{\exp[\frac{-(z-z')^2}{4D(t-t')}]+\exp[\frac{-(z+z')^2}{4D(t-t')}]\}. \qquad (2.76)$$

In principle one always has the choice of either representing a problem in terms of the free space Green's function and accommodating the boundary conditions by a surface integral, or using a Green's function specialized for the particular geometry and thus eliminating the surface integral. The shape of the boundary and the nature of the boundary condition will, of course, influence the choice. For the problem at hand the second representation is the simpler one.

Since impurity predeposition starts with zero initial concentration and has sources only at the surface inside the window, the integral (2.76) simplifies further to

$$C(\vec{r},t) = \int_0^t\int_{-\infty}^{\infty} dz\int_{W'} G(\vec{r},t|\vec{r}',t')S(x',y',t')\delta(z')dx'dy'dt'$$

$$= \int_0^t\int_{W'} 2[4\pi D(t-t')]^{-3/2}\exp[\frac{-(x-x')^2-(y-y')^2-z^2}{4D(t-t')}]\cdot$$

$$\cdot S(x',y',t')dx'dy'dt', \qquad (2.77)$$

W' indicating the area spanned by the window. For all field points inside the window the boundary condition requires that

$$C(\vec{r}_w,t) = C_s = \int_0^t \int_{W'} 2[4\pi D(t-t')]^{-3/2} \exp\left[\frac{-(x_w-x')^2-(y_w-y')^2}{4D(t-t')}\right] \cdot S(x',y',t')dx'dy'dt', \quad (2,78)$$

This equation defines the source distribution; $S(x,y,t)$ ought to be determined from it and inserted into (2,77), from which the profile $C(\vec{r},t)$ could then be evaluated by integration.

At first glance the diffusion problem looks even more formidable in this integral representation than it did in the differential formulation. It cannot be solved formally in closed form. The method of solution to be developed is specialized and heuristic, rather than general and formal and yields only an approximation. It uses analytic approaches as far as practicable before resorting to numerical techniques.

Since the integrals over time in Equations (2,77) and (2,78) are identical, a considerable simplification of the problem could be realized if they were solvable in closed form. This is indeed possible, if the source distribution is approximated by a sum of terms of the form $S_i(x,y)T_i(t)$, with T_i chosen such that it is integrable in Equations (2,77) and (2,78). The search for such functions and their integrals proves unrewarding, and the integrals in the sparse examples lead to Mathieu and Whittacker functions. A more elegant and efficient way to solve this part of the problem exists, however.

Consider a time dependent point source of strength $T_i(t)$, $t > 0$, located at $\vec{r}'$. This source generates a profile with spherical symmetry with respect to $\vec{r}'$, given by the integral

$$G_i(\vec{r},t|\vec{r}') = \int_0^t 2[4\pi D(t-t')^{-3/2}\exp[\frac{-(\vec{r}-\vec{r}')^2}{4D(t-t')}]T_i(t')dt', \qquad (2.79)$$

G_i being the spatial Green's function associated with the particular source function T_i. The Green's function can be written formally in terms of transformed coordinates, $\vec{R} = \vec{r} - \vec{r}'$, as $G_i(\vec{R},t) = G_i(R,t)$, and it satisfies the radial part of the diffusion equation

$$\partial G_i/\partial t = D[\partial^2 G_i/\partial R^2 + (2/R)\partial G_i/\partial R]. \qquad (2.80)$$

As is well known, the substitution

$$G_i = C_i/R \qquad (2.81)$$

leads to the condition

$$\partial C_i/\partial t = D\partial^2 C_i/\partial R, \qquad (2.82)$$

i.e., the function $C_i(R,t)$ solves the one-dimensional diffusion equation in the argument R. Thus any one-dimensional profile can be transformed into a Green's function of the type (2.79).

The source function associated with such a Green's function is simply equal to the total flux which emanates from the source point,

$$T_i(t) = \lim_{R\to 0}[-4\pi R^2 D\partial G_i(R,t)/\partial R] = 4\pi D C_i(0,t). \qquad (2.83)$$

Hence all one-dimensional profiles of the sets given by the expressions (2.16) and (2.17) transform at once into spatial Green's functions and associated source functions by means of Equations (2.81), (2.82), and (2.83).

The integral (2.78) can now be approximated by a sum,

$$C_s = \sum_i \int_{W'} |\vec{r}_w - \vec{r}'|^{-1} C_i(\vec{r}_w - \vec{r}', t) S_i(\vec{r}', t) d\vec{r}', \qquad (2.84)$$

where the C_i are appropriate functions from the sets (2.16) and (2.17), and where the S_i need to be determined such that the equation is satisfied as closely as possible for all points $\vec{r}_w$ in the area spanned by the window.

Consider the window of infinite dimensions as an illustration. The solution is the one-dimensional profile $C(z,t) = C_s \operatorname{erfc}(z/\sqrt{4Dt})$ and the flux at the surface equals $C_s D/\sqrt{\pi Dt}$. In an image system equal fluxes emanate from the sources in the xy-plane into the half spaces $z > 0$ and $z < 0$, and thus the source strength is equal to $2C_s D/\sqrt{\pi Dt}$. A comparison of this source function with Equations (2.16), in view of (2.83), indicates that the profile (2.16.0) yields the correct time dependence of the source function. The decomposition of the product $2C_s D/\sqrt{\pi Dt}$ into the factors $S(x,y)$ and $T_o(t)$ is arbitrary in principle, but it is convenient to choose that form of $T(t)$ which results from inserting (2.16.0) into (2.83),

$$T_o(t) = 4\pi D/\sqrt{\pi t}; \quad S(x,y) = C_s/(2\pi\sqrt{D}). \qquad (2.85)$$

Under these conditions the spatial Green's function assumes the form

$$G_o(\vec{r},t|\vec{r}') = \pi^{-\frac{1}{2}}t^{-\frac{1}{2}}|\vec{r}-\vec{r}'|^{-1}\exp[(\vec{r}-\vec{r}')/\sqrt{4Dt}]^2. \qquad (2.86)$$

The diffusion profile is obtained from the integral

$$C(z,t) = [C_s/(2\pi\sqrt{D})]\int_{z'=0} G_o(\vec{r},t|\vec{r}')dx'dy'$$

$$= \frac{C_s}{2\pi\sqrt{D}}\int_0^{2\pi}\int_0^{\infty}\pi^{-\frac{1}{2}}t^{-\frac{1}{2}}(\rho'^2+z^2)^{-\frac{1}{2}}\exp[\frac{-(\rho'^2+z^2)}{4Dt}]\rho'd\rho'd\phi', \qquad (2.87)$$

which has been written in polar coordinates for points on the z-axis. The substitution $4Dtu^2 = \rho'^2+z^2$ transforms the integral into the form

$$C(z,t) = [C_s/(2\pi)](2/\sqrt{\pi})\int_0^{2\pi}\int_{z/\sqrt{4Dt}}^{\infty}\exp-u^2du = C_s\mathrm{erfc}(z/\sqrt{4Dt}), \qquad (2.88)$$

which contains the key to a few other results which will shortly prove useful. Consider the concentration at the origin of a circular sector of sources with angle α and radius r_o. This geometry leads to upper integration limits α and $\sqrt{z^2+r_o^2}/\sqrt{4Dt}$ respectively in the integral (2.88) and hence to a concentration

$$C(\rho=0,z,t) = C_s[\alpha/(2\pi)]\{\mathrm{erfc}(z/\sqrt{4Dt})-\mathrm{erfc}[(z^2+r_o^2)^{\frac{1}{2}}/\sqrt{4Dt}]\}. \qquad (2.89)$$

In particular, the surface concentration equals

$$C(0,0,t) = C_s[\alpha/(2\pi)]\mathrm{erf}(r_o/\sqrt{4Dt}). \qquad (2.90)$$

Returning now to the problem of the predeposition diffusion through a rectangular window, Equation (2.90) allows one to estimate the surface concentration in the window, provided that the source distribution (2.85) is used as a first approximation of the true distribution.

This primary concentration is practically equal to C_s in the center region of the window and drops to 0.99 C_s at points one and a half diffusion lengths from the edges. At the edges, far away from the corners, the concentration is 0.5 C_s; and within one and a half diffusion lengths from the corners, it declines from this value to 0.25 C_s at the corners. Since the window dimensions are typically an order of magnitude larger than the diffusion length in practical cases, this first approximation is good over most of the window area. To improve upon it, secondary sources, strongest at the edges and corners of the window and tapering off toward its interior within about one and a half diffusion lengths, are necessary. These sources will depend on the time and on the location, and their determination is far from straightforward.

As long as the diffusion length is considerably smaller than the window dimensions, the neighborhood of each edge and each corner may be considered individually and approximated by one half or one quadrant of the xy-plane respectively. The latter configurations are amenable to the Boltzmann transformation [13, pp. 89-90],

$$\vec{\rho} = (\xi,\eta,\zeta) = \vec{r}/\sqrt{4Dt}, \tag{2.91}$$

which transforms the diffusion equation into the form

$$2\vec{\rho}\cdot\nabla_{\rho}C(\vec{\rho})+\nabla^{2}_{\rho}C(\vec{\rho}) = 0. \tag{2.92}$$

The boundary conditions retain their form; e.g., constant concentration in the first quadrant of the xy-plane leads to constant concentration in the first quadrant of the $\xi\eta$-plane. The profile $C(\vec{\rho})$ is a universal

solution of the problem from which the profiles in real space are obtained by replacing $\vec{\rho}$ by $\vec{r}/\sqrt{4Dt}$. The gradients in real space and in $\vec{\rho}$-space are related by the equation

$$\nabla_{\vec{r}} = (1/\sqrt{4Dt})\nabla_{\vec{\rho}}, \tag{2,93}$$

and the flux in real space can be written as

$$\vec{F} = -D\nabla_{\vec{r}}C(\vec{r},t) = -\sqrt{D/(4t)}\nabla_{\vec{\rho}}C(\vec{\rho}). \tag{2.94}$$

The strengths of sources, being proportional to the magnitude of fluxes, scale in the same way. These results lead to a valuable insight into the evolution of the real sources with time. Since the sources in real space are given in terms of the sources $S_{\vec{\rho}}(\vec{\rho})$ by

$$S(\vec{r},t) = \sqrt{D/(4t)}S_{\vec{\rho}}(\vec{r}/\sqrt{4Dt}), \tag{2,95}$$

where $\vec{r}$ is the vectorial distance of the source from the corner, the sources near the corner are related by

$$S(\vec{r},t) = \alpha S(\alpha\vec{r},\alpha^2 t). \tag{2,96}$$

By the same token, sources near the edge of the window, but far away from the corners, are related by

$$S(h,t) = \alpha S(\alpha h,\alpha^2 t), \tag{2,97}$$

where h denotes the distance from the edge.

The primary sources satisfy these conditions because they are independent of $\vec{r}$ and proportional to $t^{-\frac{1}{2}}$. The condition (2,97) also imposes functional relationships between all secondary sources, such that the time evolution of one of them is representative of all the

others. To find this time function, consider a source far from the corners but only a small distance, h, inside an edge. Initially, i.e., when h for a particular point is much larger than the diffusion length, the primary sources alone suffice to maintain the surface concentration at C_s; the secondary source starts with zero strength. Much later, when h has become much smaller than the diffusion length, the secondary source will dominate and will have to support nearly stationary flow conditions; its strength will be nearly constant. The simplest approximation to this behavior is a source which initially grows proportionally with the time and after an interval T continues with constant strength, reached at t = T. Such a time function is easily constructed as a superposition of two sources, linear in time, with equal strength and opposite sign, one delayed by T. The linear source and its Green's function result from the function (2,17,2). Equation (2,83) suggests the numerical constants, while the condition (2,97) and dimensional analysis demand the dimensioned parameters of the source function to be in the form

$$S(h,t) = 4\pi a C_s Dt/(hT); \quad T = bh^2/(4D), \tag{2.98}$$

where a and b are two free, dimensionless constants to be determined, and where T is the interval of linear rise. To prove the adequacy of the functional form of the secondary sources, and to determine the values of a and b, it is necessary to calculate the surface concentration in the window by means of Equation (2,84). This can only be done by numerical computation.

To this end the surface of a square window has been subdivided into 10,000 square elements within each of which the source distribution has been assumed constant in space. The source distribution in each element has then been replaced by a point source of equivalent strength at the center. Thereupon the concentration at these points due to the point sources has been calculated. Since the self term of a point source--the concentration at the source point itself--is infinite, the self terms must be calculated by integration of the source distribution over the elemental area. The integration of the primary sources has been demonstrated for a circular area. The function (2.90), with an effective radius, r_o, equal to 0.6 times the dimension of the elemental area has been used as the primary self term. The spatial Green's function of the secondary source, Equation (2.98), is

$$G_1(\vec{r},t|\vec{r}') = [4aC_s t/(hT|\vec{r}-\vec{r}'|)]i^2\mathrm{erfc}[(\vec{r}-\vec{r}')^2/(4Dt)]^{\frac{1}{2}}, \qquad (2.99)$$

and its integration over a circular surface element around the field point proceeds analogously to the steps (2.87) and (2.88), resulting in the functional form

$$C_s(0,0,t) = [16\pi aC_s t\sqrt{Dt}/(hT)][i^3\mathrm{erfc}(0)-i^3\mathrm{erfc}(r_o/\sqrt{4Dt})], \qquad (2.100)$$

The computation procedure uses four secondary sources at each lattice point, according to the distances to the four edges. This approximation automatically leads to an enhancement of the source strength toward the corners, although the functional form is not quite correct. One might think of more sophisticated methods of adjusting the secondary sources

near the corners or of determining the self terms, but the improvement would at best be marginal. The most critical inaccuracy lies in the discretization of the source distribution into point sources, a problem which Fourier [20, p. 377] pointed out. However, in computing the concentration due to the primary sources only, at a point sufficiently far away from the edges, one finds that the discretization error declines to less than 3% as the diffusion length grows beyond five times the element size. Under these conditions the combination of all sources, using the values a = 0.0224 and b = 36, produced a surface concentration profile in the window which deviates less than 2% from C_s over four fifths of the area and deviates at no point more than 10%, until the diffusion length reaches one third of the width of the window. At this upper limit it abruptly starts increasing. This time region, however, is of little interest for the problems under consideration.

In summary, a source distribution has been found, which describes the predeposition diffusion through a rectangular window quite well. Since the largest inaccuracies in the profile are closest to the sources, and since the profile due to each individual source is known in closed form and therefore with perfect accuracy, the source distribution allows the determination of the profile with at most a few per cent error at any point in space. Although the calculation of the concentration at one field point entails the accumulation of 90,000 individual contributions, each obtained by the evaluation of transcendental functions, this procedure is still vastly more efficient than an iterative scheme, because the profile is computed only for the

field points and times of interest, while the iterative schemes require the many times recurring calculation of the profile on a fine lattice of points over a large volume. In spite of the simplicity of the individual iterations, the vast number of time steps and lattice points, which are of little interest for the final solution, lead to much larger computational expense, not to mention the problems with accuracy in the regions of low concentration.

The source distribution so far models only the predeposition diffusion. To describe a two step diffusion, one only needs to turn off all sources at the end of the predeposition cycle. Step functions and linear time functions suffice to compensate for the secondary sources, while the primary sources can be cancelled by the techniques outlined in Section 2.3. However, two step diffusions are hardly used in connection with very small windows.

2.5 Three-Dimensional Impurity Profiles from Rectangular Diffusion Windows--Approximate Analytic Solutions

The problem of calculating the three-dimensional impurity profile due to the predeposition diffusion through a rectangular window has essentially been solved in the previous section. However, the result of the numerical calculation will always be an array of numbers, which is ill suited for deriving relationships between structural parameters and the characteristics of the solution. Therefore, this section is devoted to approximate analytic solutions of the same problem and their comparison with the numerically determined profile. The emphasis will be more on the simplicity and tractability of the approximate formulations than on accuracy. An approximation of

simple form will eventually be helpful in relating the geometrical and physical parameters of the device to its electrical characteristics.

An approximate solution should at least produce a constant surface concentration over most of the window surface, and the vertical profile far inside the window should be correct. If the solution also satisfies the diffusion equation and scales with the diffusion length, it is universal with respect to the time variable and the analysis of the approximation error for one point in time suffices. Again, the restriction shall hold that the window dimensions are larger than three diffusion lengths.

The execution of the divergence and time derivative operations on a product of the form

$$C(x,y,z,t) = C_1(x,t)C_2(y,t)C_3(z,t) \tag{2.101}$$

where the C_i are solutions of the one-dimensional diffusion equation, shows at once that such products, and hence any linear combination thereof, describe a valid three-dimensional diffusion profile. The simplest such approximation, which reflects the guidelines mentioned above, for a window $|x| < a$, $|y| < b$, is

$$C(x,y,z,t) = \tfrac{1}{4}C_s \operatorname{erfc}(z/\sqrt{4Dt})\{\operatorname{erf}[x+a)/\sqrt{4Dt}]-\operatorname{erf}[(x-a)/\sqrt{4Dt}]\}\cdot$$

$$\cdot\{\operatorname{erf}[(y+b)/\sqrt{4Dt}]-\operatorname{erf}[(y-b)/\sqrt{4Dt}]\}. \tag{2.102}$$

The deficiencies of this approximation are immediately perceptible: Inside the window the concentration declines toward the edges; it equals $C_s/2$ at the edges, far away from the corners, and $C_s/4$ at the corners. Outside the window the flux through the surface does not

vanish. Figures 2.4 and 2.5 give a more comprehensive picture of the overall quality of the approximation. Figure 2.4 shows contours of constant concentration in the xy-plane around a corner of the window. For comparison the contours derived from the numerical solution are indicated by broken lines. Figure 2.5 illustrates the profiles in a vertical plane perpendicular to the edge of the window and in the vertical plane of symmetry through a corner of the window. Again, contours of constant concentration are shown with solid lines for the approximation and with broken lines for the numerical solution. The important features of the profile for the electrical behavior of the structure are the location and curvature of the metallurgical junction and the distance of outdiffusion, i.e., the location and characteristics of a particular contour of constant concentration between about $10^{-3}C_s$ and $10^{-5}C_s$. The gradient of the concentration near the junction is also important. Figure 2.5 indicates that the simple approximation describes the essential features of the profile quite closely in the region of interest and at a depth equal to or greater than the distance from the window edge outward. The approximation yields an outdiffusion at the surface which is about 10% too large.

Figure 2.5 would typically describe the emitter diffusion profile for a bipolar device. In this case the deeper regions have more influence on the electrical characteristics than the regions near the surface--the approximation can be considered sufficiently good. On the other hand, the electrical characteristics of insulated gate field effect transistors (IGFET) depend most strongly on the impurity profile near the surface, and the approximation is insufficient for accurate

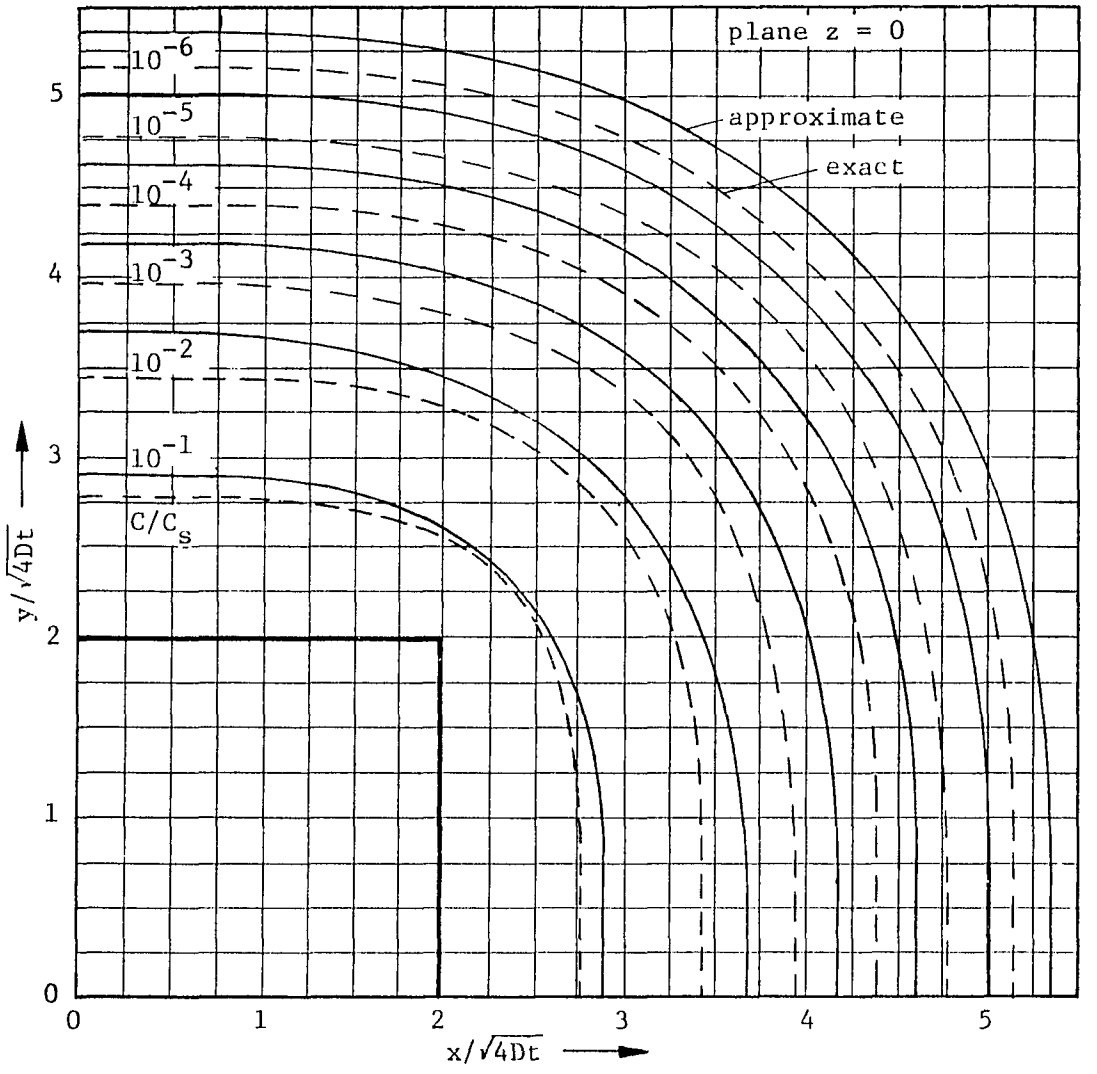

Figure 2.4, Contours of Constant Dopant Concentration in One Quadrant of the Surface Around a Square Window

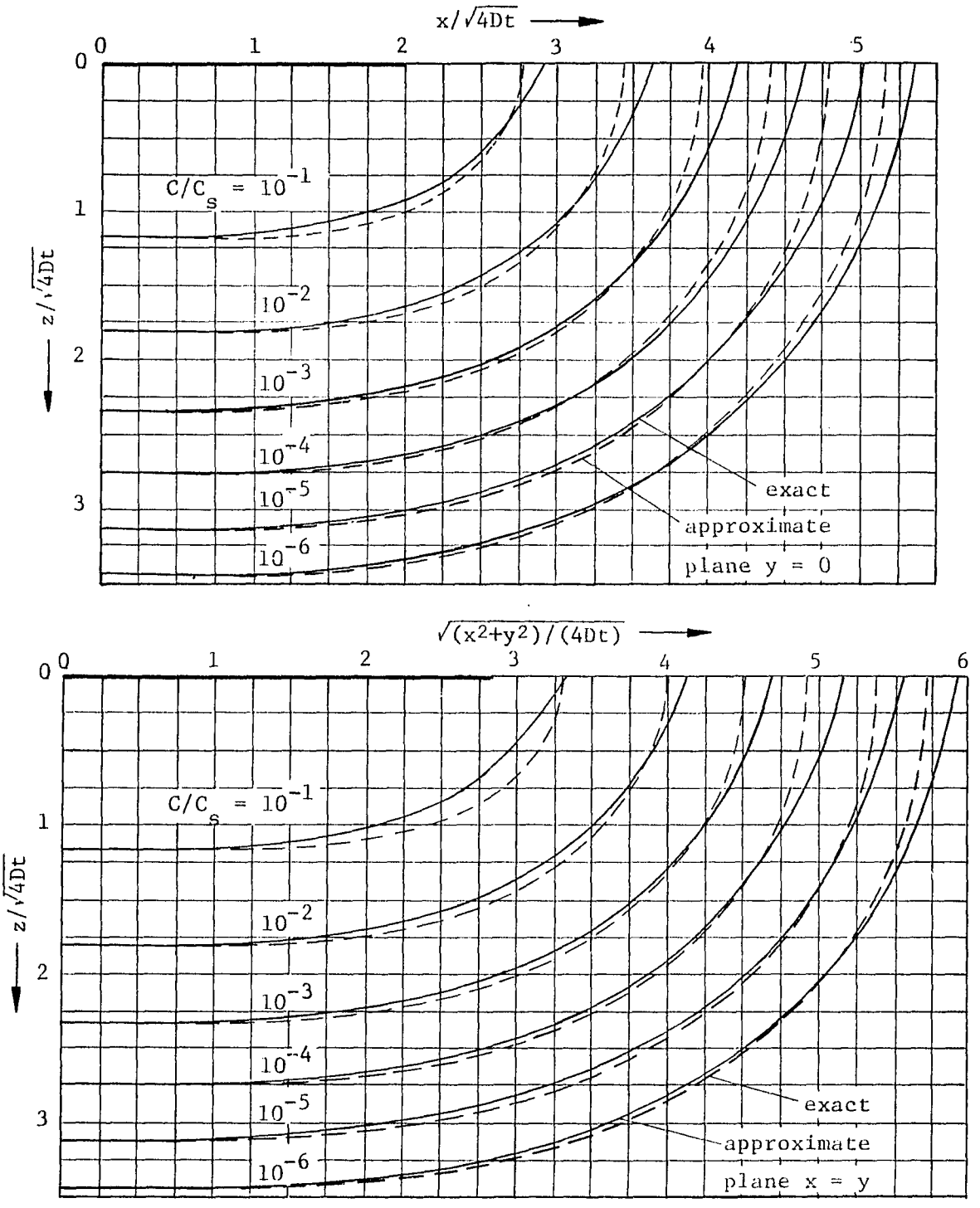

Figure 2.5. Contours of Constant Dopant Concentration in Two Planes of Symmetry Normal to the Square Window

modeling of such devices. Since accurate models of IGFETs require the solution of the nonlinear carrier transport equations, which fall outside the realm of good analytic approximations, so far, one may for these cases use numerically calculated impurity profiles also.

Further refinements of the approximation would result from the addition of other products of functions from the sets (2.16) and (2.17), notably one whose x- and y-dependent parts change sign at the edges of the window. No new insights into the approximation technique can be gained, however, by pursuing further refinements.

2.6 Diffusion with Concentration Dependent Diffusivity

The impurity profiles derived so far in this chapter stem from linear diffusion processes with constant diffusivity. Many practical diffusions occur at sufficiently high concentrations to make the process nonlinear by virtue of a concentration dependent diffusivity.

The concentration dependence of the diffusivity is a consequence of the physical constraint that no electric currents flow in the specimen. To satisfy this condition an electric field inside the specimen builds up, and this field is proportional to the impurity concentration gradient [21]. Thus, the diffusion component as well as the drift component of the flux of ionized impurities depend on the concentration gradient, and an effective diffusivity can be defined which assumes the form

$$D(C) = D_o\{1+[1+(2n_i/C)^2]^{-\frac{1}{2}}\}, \qquad (2.103)$$

where D_o is the diffusion constant for impurity concentrations which are much lower than the intrinsic carrier concentration, n_i.

Lehovec and Slobodskoy [22] have published a semi-analytic method for calculating one-dimensional impurity profiles from non-linear predeposition diffusion processes, based on the effective diffusivity (2.103). They demonstrate that these profiles have a low concentration tail described by a complementary error function of the form

$$C(z,t_p) = C_s' \operatorname{erfc}(z/\sqrt{4D_o t_p}) \ll n_i, \tag{2.104}$$

where C_s' is a virtual surface concentration which depends on C_s/n_i.[1] If C_s itself is much larger than n_i, then there is a high concentration region near the surface, given by

$$C(z,t_p) = C_s \operatorname{erfc}(z/\sqrt{8D_o t_p}) \gg n_i. \tag{2.105}$$

The authors give the relationship between C_s'/n_i and C_s/n_i in a graph. With these tools it is simple to determine one-dimensional nonlinear predeposition profiles, except in the transition region, which encompasses a region of about one order of magnitude in concentration and about one diffusion length in depth, centered about the intersection of the curves (2.104) and (2.105).

A somewhat different approach, while not leading to a tractable way for actually calculating nonlinear profiles, produces

---

1. Note that n_i is the intrinsic carrier concentration at the diffusion temperature.

relationships between the profiles from linear and nonlinear diffusion through the same window geometry.

In Section 2.3 the Kirchhoff transformation was introduced to relate the nonlinear diffusion Equation (2.4): $\nabla\cdot(D\nabla C) = \partial C/\partial t$, to the simpler form (2.6): $D_o f(C)\nabla^2 B = \partial B/\partial t$. Writing the diffusivity as in (2.3): $D = D(C) = D_o f(C)$, the relation between B and C is given by (2.5): $B(C) = \int f(C')dC'$. For the form of the diffusivity given in (2.103), closed form relationships between B, C, and f exist;

$$B(C) = n_i\{(C/n_i)+[4+(C/n_i)^2]^{\frac{1}{2}}-2\}, \qquad (2.106)$$

$$C(B) = \tfrac{1}{2}B(4+B/n_i)/(2+B/n_i), \qquad (2.107)$$

$$f(B) = 1+\{1+[(8n_i^2+4Bn_i)/(4Bn_i+B^2)]^2\}^{\frac{1}{2}}. \qquad (2.108)$$

Figure 2.6 shows plots of $f(C) = D/D_o$ and of B/C as functions of C/n_i and illustrates two regions where f and B/C are constant, as well as the transition regions in between. Since the mapping from B to C and vice versa is unique, knowledge of one quantity implies knowledge of the other.

Now, write Equation (2.6) in the form

$$D_o f[B(\vec{r},t)]\nabla^2 B(\vec{r},t)-\partial B(\vec{r},t)/\partial t = 0 \qquad (2.109)$$

and consider a transformation

$$\partial\tau/\partial t = D_o f[B(\vec{r},t)]; \quad \tau(t=0) = 0 \qquad (2.110)$$

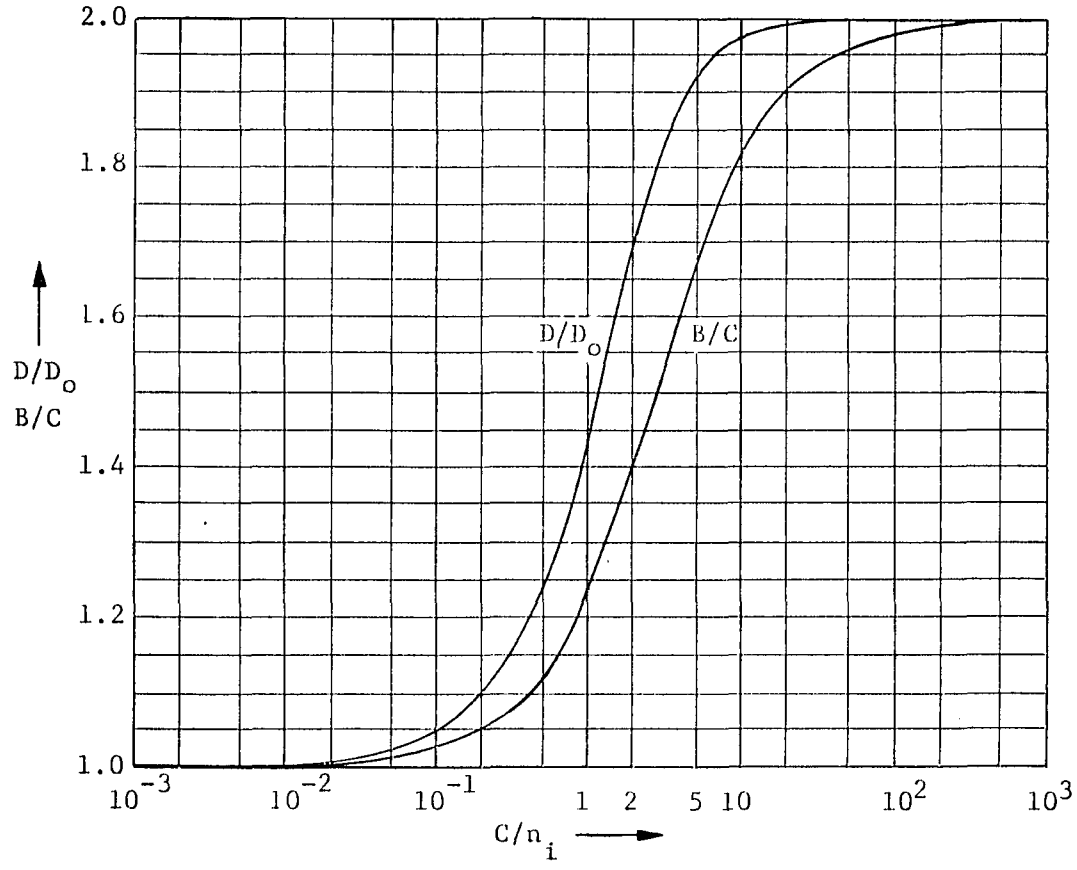

Figure 2.6. The Normalized Diffusivity, D/D_o, and Its Normalized Integral Over Concentration, B/C, as Functions of C/n_i

and its integral form

$$\tau(\vec{r},t) = D_o\int_0^t f[B(\vec{r},t')]dt'. \tag{2.111}$$

Since C is positive and B is a monotonic positive function of C, $\tau(r,t)$ is a single valued function of t. In fact, the mapping is unique from τ to t also. Thus, one can formally express B as a function of τ also. Rewriting Equation (2.109) in terms of $B(\vec{r},\tau)$ requires the application of the chain rule to the time derivative,

$$\partial/\partial t = (\partial\tau/\partial t)\partial/\partial\tau = D_o f[B(\vec{r},t)]\partial/\partial\tau, \tag{2.112}$$

and, upon division by $D_o f[B(\vec{r},t)]$, leads to

$$\nabla^2 B(\vec{r},\tau)-\partial B(\vec{r},\tau)/\partial\tau = 0. \tag{2.113}$$

Therefore $B(\vec{r},\tau)$ is a solution of the linear diffusion equation. Since the boundary condition on B in the window is simply $B(\vec{r}_w) = B(C_s)$ by virtue of (2.106), $B(\vec{r},\tau)$ will be described by one of the linear predeposition profiles of Sections 2.3, 2.4, and 2.5. The remaining problem is the transformation of $B(\vec{r},\tau)$ into $C(\vec{r},t)$.

Formally, the scalar field $B(\vec{r},\tau)$ leads to a scalar field $f[B(\vec{r},\tau)]$ by virtue of Equation (2.108). With the aid of (2.110) one then obtains

$$t(\vec{r},\tau) = (1/D_o)\int_0^\tau d\tau'/\{f[B(\vec{r},\tau')]\} \tag{2.114}$$

as a unique relationship between t and τ for each point of the spatial region. The solution $B(\vec{r},t)$ is a cross-section through the solution

$B(\vec{r},\tau)$ such that t is the same at all points in the diffusion region, rather than τ. The function $B(\vec{r},t)$, in turn, transforms uniquely into $C(\vec{r},t)$ by virtue of Equation (2.107).

As an example consider the three-dimensional profile due to a diffusion through a rectangular window with a surface concentration $C_s >> n_i$. As outlined, the profile is characterized by regions of high and low concentrations with respect to n_i, which are separated by a transition region. The low concentration region is characterized by

$$C << n_i; \quad f = 1; \quad C = B; \quad \tau = D_o t. \tag{2.115}$$

The profiles of the linear and nonlinear diffusion process differ by at most a multiplicative constant besides the usual scaling with the diffusion length.

In the transition region the relationships

$$C \simeq n_i; \quad 1 < f < 2; \quad C < B < 2C; \quad D_o t < \tau < 2D_o t \tag{2.116}$$

apply. Not much can be said about the details of the profile in this region because the relationship between t and τ need not be the same on a constant concentration surface of the linear profile. Therefore the concentration of the nonlinear profile need not be constant on this surface.

In the region of high concentration, finally, the transformation (2.114) has to be applied at all points because the concentration starts from zero everywhere in this region and reaches the high values only as the profile expands with time. Nevertheless, the farther away from the transition region a point is, the larger is the fraction of

time during which the concentration is high, and the closer will the integral approach the value $\tau = 2D_o t$. Hence, the conditions for this region are

$$C >> n_i; \quad f = 2; \quad C = B/2; \quad \tau = 2D_o t \tag{2.117}$$

to a very close approximation. In this region the linear profile with $D = 2D_o$ represents the nonlinear profile.

For the one-dimensional predeposition profiles these considerations lead to the same result as the calculations of Lehovec and Slobodskoy. For the three-dimensional profiles from diffusion through a window, the development ensures that the surfaces of constant concentration of the nonlinear profile in the regions of high and of low concentration are identical to those of linear profiles of appropriately chosen diffusion length. Since all these surfaces approach planes underneath the window surface and are spaced according to the one-dimensional solution, the concentration of each of the surfaces can be determined from the one-dimensional profile calculated by Lehovec and Slobodskoy. In the transition region, surfaces of the linear and of the nonlinear profile need not be congruent, but they can be expected to deviate only slightly from each other. When the metallurgical junction falls into the low concentration region, it is usually sufficient to determine the profile in this region only, and this can be done by calculating a linear profile with the appropriate effective surface concentration as given by Lehovec and Slobodskoy. It might be noted that the effective surface concentration can be up to two orders of magnitude higher than the actual surface concentration,

such that the profile may actually be required over up to eight decades of the concentration. With the methods outlined in this chapter the range of the profile can be extended to these low concentrations without loss of accuracy.

CHAPTER 3

CARRIER STATISTICS, CARRIER PROFILES, AND CARRIER TRANSPORT

3.1 Outline of the Chapter

Carrier transport phenomena form the bridge between the physical structure of solid state devices and their electrical characteristics. The following sections will thus set the stage for detailed modeling of critical regions in device structures and of entire devices. One-dimensional structures, and even homogeneous regions, will again play an important role in developing the methods for the analysis of multi-dimensional effects.

The improvement in accuracy expected from the inclusion of multidimensional effects in the models also warrants a more accurate determination of the carrier statistics at high concentrations, than the generally used simplified semiconductor theories can yield.

The behavior of the carriers is governed by the laws of thermodynamics. The development of the transport equations from the theory of irreversible thermodynamics provides the best insight into the often tacit assumptions which underly the conventional carrier transport theories. Thermodynamics also unveils the relationships between the transport coefficients and can include electrothermal and thermoelectric effects from the start and in a natural way. Last, but not least, the theory formulates the transport processes in terms of a variational principle. This opens up the possibility of numerically

calculating the carrier and energy fluxes by means of the finite element method. The shortcomings which made this method unsuitable for the computation of impurity profiles are much less critical here, because one is mainly interested in the total currents, with little or no interest in regions of extremely low flux or carrier concentration. However, irreversible thermodynamics is only applicable to small deviations from equilibrium.

The next section briefly outlines the theory of irreversible thermodynamics of carrier transport. The formulation highlights the determining influence of the carrier distributions at thermostatic equilibrium on the transport properties of the material. This result justifies the devotion of a large part of the chapter to the study of the equilibrium distributions of the carriers.

The third section reviews the essential features of the equilibrium statistics of carriers in homogeneous material in thermostatic equilibrium, covering the ranges of impurity concentrations and temperatures normally used in integrated circuit technology. Approximations are derived which allow the treatment of carrier concentrations up to degeneracy with formulations almost as simple as those of the traditional simplified semiconductor theories.

The fourth and fifth sections are devoted to the determination of carrier distributions in nonuniformly doped material in thermostatic equilibrium. One would intuitively expect that, with the exception of the regions near metallurgical junctions, or other abrupt changes of the impurity concentration, the local majority carrier concentration would be related to the local impurity concentration in almost the same

way as in homogeneous material. The derivations of the fourth section confirm this view and provide quantitative relationships which are very accurate for concentrations above about 10^{16} cm^{-3}. The remaining regions are exclusively those near metallurgical junctions, surfaces, and contacts.

The fifth section demonstrates that junctions encountered in practice at thermostatic equilibrium are always characterized by a transition region which is depleted of free carriers. The space charge of this region is essentially that of the ionized impurity atoms. The "depletion approximation"--neglecting the free carriers altogether--is a good approximation for these regions.

The sixth section is devoted to a partly qualitative, partly quantitative treatment of the phenomena at the interfaces of the semiconductor with contact metals and with insulators. The potentials at the contacts and the processes at the surfaces form the boundary conditions for the transport problems which need to be solved in order to obtain the device characteristics. The first part of the section highlights the few important parameters which determine the characteristics of the interfaces in thermostatic equilibrium: work function, electron affinity, surface charge density, interface dipole layer, and space charge layer in the semiconductor. The second part of the section considers the behavior of the interfaces under nonequilibrium conditions.

3.2 Thermodynamics of Carrier Transport in Semiconductors

Smith, Janak, and Adler [23] have written an excellent treatise of the thermodynamic theory of heat and carrier transport in solids and

of related topics. This section is not intended to rederive or re-justify these theories; the pertinent equations will be introduced mainly for defining salient quantities. A few thoughts useful to the modeling of semiconductor devices will be added.

It is important to keep in mind that the theory of thermodynamics concerns itself with macroscopic quantities which are statistical averages over sufficiently large ensembles. For example, a differential, dn, of the electron concentration must represent at least hundreds of electrons, such that the statistical fluctuations from increment to increment are negligible. When this point of view is untenable, because the unit volume is too small, one may alternatively consider averages over an ensemble of identical structures. In these cases one may find that the electrical characteristics of the devices exhibit random variations due to statistical fluctuations [24].

A semiconductor structure in an environment of uniform temperature which is disconnected from any power sources is in thermostatic equilibrium. No fluxes of energy or particles transcend it, and it is in a state of maximum entropy. The theory of reversible thermodynamics is applicable to this state. A device in operation, on the other hand, experiences fluxes of energy and carriers and is out of thermostatic equilibrium. If the operating conditions are time-independent the device is said to be in a steady state, or in thermodynamic equilibrium. If the deviations from thermostatic equilibrium are not too large, the system can be described by the theory of irreversible thermodynamics. This theory postulates that the flow processes adjust themselves such that the entropy production is

minimized. More will be said about the loose description, "not too large," from time to time.

An incremental change in entropy per unit volume in a system of acceptors, donors, holes and electrons in a semiconductor is given as

$$ds = \delta q/T = (1/T)du-(\mu_n/T)dn-(\mu_p/T)dp-(\mu_D/T)dD-(\mu_A/T)dA, \qquad (3.1)$$

where T is the absolute temperature; μ_n, μ_p, μ_D and μ_A are respectively the electrochemical potentials of electrons, holes, neutral donors (electrons on donor atoms) and neutral acceptors; where $(1/T)$, (μ_n/T), (μ_p/T), (μ_D/T) and (μ_A/T) are the thermodynamic potentials or intensive variables; and where ds, du, dn, dp, dD and dA are, respectively, differentials of the entropy density, the energy density, the density of electrons, the density of holes, the density of neutral donors, and the density of neutral acceptors. The latter variables are called extensive variables. Phenomenologically the entropy in a volume element can be changed by generating carrier pairs, by capturing carriers to neutralize ionized impurities, and by net fluxes of energy or carriers through the boundary surface of the volume element. The entropy rates and fluxes are the product of the rates and fluxes of the extensive variables with their conjugate intensive variables. The entropy generation rate per unit volume is

$$g_s = \nabla\cdot[(1/T)\vec{j}_u-(\mu_n/T)\vec{j}_n-(\mu_p/T)\vec{j}_p]+$$

$$+(1/T)\dot{u}-(\mu_n/T)\dot{n}-(\mu_p/T)\dot{p}-(\mu_D/T)\dot{D}-(\mu_A/T)\dot{A}, \qquad (3.2)$$

the dot denoting partial differentiation with respect to time. The fluxes and rates satisfy the continuity equations,

$$\nabla\cdot\vec{j}_u+\dot{u} = 0, \tag{3.3}$$

$$\nabla\cdot\vec{j}_n+\dot{n} = g_n, \tag{3.4}$$

$$\nabla\cdot\vec{j}_p+\dot{p} = g_p, \tag{3.5}$$

$$\dot{D} = g_D, \tag{3.6}$$

$$\dot{A} = g_A, \tag{3.7}$$

where $\vec{j}$ and g denote fluxes and generation rates respectively. Since electron transitions above the lowest order conduction band are highly unlikely, the generation rates are constrained by

$$g_n+g_D = g_p+g_A. \tag{3.8}$$

The maximization of the entropy at thermostatic equilibrium requires that

$$\mu_{no} = \mu_{Do} = -\mu_{po} = -\mu_{Ao}. \tag{3.9}$$

Subtracting these equilibrium values from the electrochemical potentials, i.e., writing

$$\delta\mu_i = \mu_i-\mu_{io} \tag{3.10}$$

and inserting Equations (3.3) to (3.10) into (3.2) results in

$$g_s = \vec{j}_u\cdot\nabla(1/T)+\vec{j}_n\cdot\nabla(-\mu_n/T)+\vec{j}_p\cdot\nabla(-\mu_p/T)-$$

$$-g_n\delta\mu_n/T-g_p\delta\mu_p/T-g_D\delta\mu_D/T-g_A\delta\mu_A/T. \tag{3.11}$$

The theory of irreversible thermodynamics now postulates that the gradients of the thermodynamic potentials are the driving forces for

the fluxes, and that their deviations, $\delta\mu_i/T$, from the equilibrium values are the driving forces for the generation rates. It is also postulated that the fluxes and rates are linear functions of the respective driving forces. The linearity is expressed by the equations

$$\begin{pmatrix} \vec{j}_n \\ \vec{j}_p \\ \vec{j}_u \end{pmatrix} = \begin{pmatrix} L_{nn} & L_{np} & L_{nu} \\ L_{pn} & L_{pp} & L_{pu} \\ L_{un} & L_{up} & L_{uu} \end{pmatrix} \begin{pmatrix} \nabla(-\mu_n/T) \\ \nabla(-\mu_p/T) \\ \nabla(1/T) \end{pmatrix}. \tag{3.12}$$

The transport coefficients, L_{ij}, are in general second rank tensors, but the symmetry of cubic crystals like silicon and germanium reduces them to scalar quantities. The theory also states that the matrix of transport coefficients in (3.12) is symmetric. Partially executing the gradient operation results in

$$\begin{pmatrix} \vec{j}_n \\ \vec{j}_p \\ \vec{j}_u \end{pmatrix} = \begin{pmatrix} L_{nn}/T & L_{np}/T & (L_{nn}\mu_n+L_{np}\mu_p-L_{nu})/T^2 \\ L_{np}/T & L_{pp}/T & (L_{np}\mu_n+L_{pp}\mu_p-L_{pu})/T^2 \\ L_{nu}/T & L_{pu}/T & (L_{nu}\mu_n+L_{pu}\mu_p-L_{uu})/T^2 \end{pmatrix} \begin{pmatrix} \nabla(-\mu_n) \\ \nabla(-\mu_p) \\ \nabla T \end{pmatrix}. \tag{3.13}$$

This system of equations yields the conditions for the thermostatic equilibrium: to make the fluxes vanish, the driving forces must vanish;

$$\nabla\mu_{no} = \nabla\mu_{po} = 0; \quad \nabla T = 0, \tag{3.14}$$

i.e., the electrochemical potentials and the temperature are spatially constant.

To semiconductor engineers the electrochemical potentials are better known by their equivalents, the quasi-Fermi levels or imrefs.

Their equilibrium values, μ_{no} and $-\mu_{po}$, correspond to the Fermi level. As long as the Boltzmann approximation is valid the free carriers behave like ideal gases, and the chemical potentials are logarithmic functions of the carrier densities. The electrochemical potentials are sums of the chemical potentials and the electrical potential. Both potentials are only defined to within an additive constant. The partitions used here are

$$\mu_n = \mu_{cn} + (-e)\phi = (kT)\ln(n/n_i) - e\phi, \tag{3.15}$$

$$\mu_p = \mu_{cp} + e\phi = (kT)\ln(p/n_i) + e\phi. \tag{3.16}$$

Figure 3.1 illustrates the relationships between the carrier concentrations and the energy levels in a volume element of an extrinsic, but nondegenerate semiconductor.

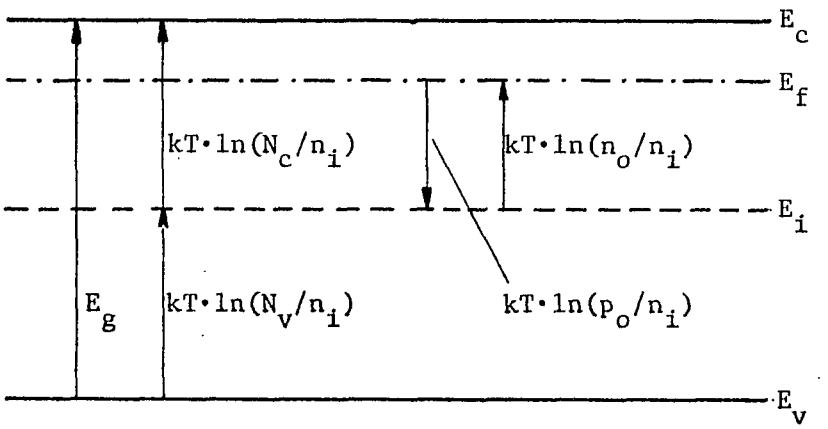

Figure 3.1. Relations Between the Energy Levels and Carrier Concentrations in an Extrinsic, Nondegenerate Semiconductor

The energies, E_c, E_i, E_f, and E_v indicate the conduction band edge, the intrinsic level, the Fermi level, and the valence band edge respectively; n_i is the carrier concentration of the intrinsic material; N_v and N_c are the effective densities of states in the valence and conduction band; and n_o and p_o are the spatially varying electron and hole concentrations at thermostatic equilibrium. Since E_f is constant throughout any structure, it is reasonable to use it as the origin for the electron and hole energy scales. The electron energy increases upwards, toward E_c, and the hole energy increases downwards. Thus, in view of (3.15) and (3.16),

$$\mu_{no} = -\mu_{po} = E_f = 0, \tag{3.17}$$

$$\phi_o = (E_{io}-E_f)/e, \tag{3.18}$$

$$\mu_{cpo} = (kT)\ln(n_o/n_i) = E_f-E_{io}, \tag{3.19}$$

$$\mu_{cno} = (kT)\ln(p_o/n_i) = E_{io}-E_f. \tag{3.20}$$

The potential ϕ_o is the "built-in potential."

When a structure is forced into a nonequilibrium state by externally applied potentials, then μ_n, μ_p, and ϕ differ from their equilibrium values. A positive potential depresses the level E_i with respect to E_{io}; a potential $\mu_n > \mu_{no}$ will appear above E_f, while a potential $\mu_p > \mu_{po}$ will appear below E_f in a sketch like Figure 3.1. In traditional carrier transport theory for semiconductor devices, the electron and hole particle current densities are specified as

$$\vec{j}_n = -nm_n\vec{E}-D_n\nabla n, \tag{3.21}$$

$$\vec{j}_p = pm_p\vec{E} - D_p\nabla p, \tag{3.22}$$

where m_n and m_p denote the electron and hole mobilities; and where D_n and D_p are the respective carrier diffusivities. For nondegenerate carrier concentrations these parameters are related by the Einstein relation,

$$D_n/m_n = D_p/m_p = kT/e, \tag{3.23}$$

Since $E = -\vec{\nabla}\phi$, one can rewrite Equations (3.21) and (3.22) as

$$\vec{j}_n = nm_n[\nabla\phi - (kT/e)(1/n)\nabla n] = (nm_n/e)\nabla[e\phi - (kT)\ln(n/n_i)]$$

$$= (nm_n/e)\nabla(-\mu_n), \tag{3.24}$$

$$\vec{j}_p = pm_p[-\nabla\phi - (kT/e)(1/p)\nabla p] = (pm_p/e)\nabla[-e\phi - (kT)\ln(p/n_i)]$$

$$= (pm_p/e)\nabla(-\mu_p). \tag{3.25}$$

The comparison of these equations with (3.13) highlights the tacit assumptions behind conventional carrier transport theory. The effects of temperature gradients are disregarded and the coefficient L_{np} is assumed to be zero. The transport coefficients L_{nn} and L_{pp} in this case assume the form

$$L_{nn} = Tnm_n/e; \quad L_{pp} = Tpm_p/e, \tag{3.26}$$

The postulated linearity of the generation rates is expressed by the equations

$$\begin{pmatrix} g_n \\ g_p \\ g_D \\ g_A \end{pmatrix} = \begin{pmatrix} a_{nn} & a_{np} & a_{nD} & a_{nA} \\ a_{np} & a_{pp} & a_{pD} & a_{pA} \\ a_{nD} & a_{pD} & a_{DD} & a_{DA} \\ a_{nA} & a_{pA} & a_{DA} & a_{AA} \end{pmatrix} \begin{pmatrix} -\delta\mu_n/T \\ -\delta\mu_p/T \\ -\delta\mu_D/T \\ -\delta\mu_A/T \end{pmatrix} \tag{3.27}$$

Equation (3,8) makes these four equations dependent and imposes constraints on the rate coefficients, The three remaining independent equations can be written as

$$\begin{pmatrix} g_n \\ g_p \\ \cdot \\ g_D \end{pmatrix} = \begin{pmatrix} a_{nn} & a_{np} & a_{nD} & (a_{nn}+a_{nD}-a_{np}) \\ a_{np} & a_{pp} & a_{pD} & (a_{np}+a_{pD}-a_{pp}) \\ a_{nD} & a_{pD} & a_{DD} & (a_{nD}+a_{DD}-a_{pD}) \end{pmatrix} \begin{pmatrix} -\delta\mu_n/T \\ -\delta\mu_p/T \\ -\delta\mu_D/T \\ -\delta\mu_A/T \end{pmatrix}. \tag{3.28}$$

In nondegenerate material the impurities are almost fully ionized, and g_A and g_D vanish, Then, $g_n = g_p$, and $a_{nn} = a_{pp} = a_{np} = a$, while all other rate coefficients vanish. Under these conditions, in view of Equations (3.10) and (3.14) to (3.16),

$$g_n = g_p = (-a/T)(\delta\mu_n+\delta\mu_p) = (-ak)\ln(np/n_i^2), \tag{3.29}$$

i,e,, the generation rates are proportional to the deviations of the np-product from its value at thermostatic equilibrium, provided that the deviations are small enough to allow the expansion of the logarithm to first order, One notes that in order to restore equilibrium the generation rates must be positive if $np < n_i^2$, which requires the constant a to be positive.

By comparing the generation rates in (3.29) with those derived from the relaxation time approximation one can identify the coefficient a. The relaxation time approximation for extrinsic, nondegenerate n-material ($n >> p$) is

$$g_n = g_p = (n_o p_o - np)/(n_o \tau_p), \tag{3.30}$$

with τ_p the minority carrier life time. For small deviations from the thermostatic equilibrium, where $np/n_i^2 \simeq 1$, one can write

$$(ak)\ln(np/n_i^2) \simeq ak(np-n_i^2)/n_i^2. \tag{3.31}$$

From (3.29), (3.30), and (3.31) one finally concludes that

$$a = p_o/(k\tau_p). \tag{3.32}$$

For p-material, the constant, a, will be given by the analogous equation, in terms of n_o and τ_n. For larger deviations of the np-product from n_i^2, Equations (3.29) and (3.30) yield progressively more different generation rates. Both are approximations, however, and a more accurate numerical calculation of the generation rates as a function of the degree of nonequilibrium, based on the equations of the Shockley-Reed-Hall theory, reveals that (3.29) yields better results overall.

In concluding this section it is worthwhile to summarize the main points. The laws of carrier transport can be expressed by two sets of linear equations, namely (3.12) and (3.27) or simplifications thereof. The proportionality constants are the transport coefficients, L_{ij}, and the rate coefficients, a_{ij}, respectively.

It should be noted that the transport coefficients given by Equation (3.26) contain the carrier concentrations as factors, and thus are themselves highly nonlinear functions of the chemical potentials. Only if the deviations from thermostatic equilibrium are small enough that $n-n_o \ll n_o$ and $p-p_o \ll p_o$, can these coefficients be treated as approximately constant and independent of the chemical potentials. This is a severe restriction which almost never applies to junction regions. The mobilities contained in the transport coefficients are comparatively weak functions of the impurity concentrations and can be considered locally constant.

The transport coefficients are intrinsic material properties and must be determined experimentally from suitable structures. In practice such specialized structures can be processed together with integrated circuit chips, either as kerf structures on each chip, or as separate test sites on the same wafer.

The driving forces for the carrier fluxes and generation rates are, respectively, the gradients of the electrochemical potentials and their deviations from thermostatic equilibrium. These potentials consist of two parts; the chemical potentials which are logarithmic functions of the carrier concentrations, and the electrostatic potential. The carrier concentrations and electrostatic potentials at thermostatic equilibrium, in turn, depend solely on the impurity profiles constituting the device structure. The remainder of this chapter is primarily devoted to the presentation of approximate analytic methods for determining the carrier distributions from the impurity profiles and other relevant features of device structures.

3.3 Equilibrium Carrier Distribution in Uniformly Doped Material

Basic theories and methods for calculating the salient parameters of equilibrium carrier distributions in moderately heavily doped material are well established in the literature. These methods, for example those outlined by Sze [25, pp. 25 ff], are now veritably classic. More recently it has been recognized that high impurity concentrations lead to considerable deviations of device properties from those predicted by these theories [26].

Modern device technology frequently uses impurity concentrations up to the solid solubilities, which lie between $2 \cdot 10^{20}$ cm^{-3} and $2 \cdot 10^{21}$ cm^{-3} for phosphorus, arsenic and boron in silicon. The resulting high majority carrier concentrations, first of all, make the use of Fermi-Dirac statistics unavoidable. They also give rise to changes in the band structure; one observes band tailing and impurity band formation. The former is the development of a tail of the density of states function which penetrates into the band gap; the latter is a dispersion of the energies of the impurity states about the level at low concentration.

Parmenter [27, 28] has investigated the band structure of alloys--which include doped semiconductors--using a purely quantum mechanical model and perturbation methods. He found band tailing, but the model does not describe impurity states in the band gap. In his models the band tailing conceptually results from the perturbation of the eigenenergies, caused by coupling between the quantum states of the

ideal crystal. The coupling results from the impurities which disturb the periodicity of the lattice.

Kane [29] studied the band structure of doped semiconductors with the aid of the Thomas-Fermi approximation. In this semi-classical model the shifts of the energies result from the spatial variation of the potential, which is caused by the statistical fluctuations of the impurity concentration. He obtained a qualitatively equivalent result to Parmenter's, although the density of states is of a different functional form. Morgan [30] used the Thomas-Fermi approximation for calculating the distribution of the energies of the impurity states.

The basic concept of the latter two calculations is quite straightforward. The local density of states function is considered to be that of an unperturbed crystal, but with the energy bands pegged to the local electrostatic potential, ϕ_l, in the same way as illustrated in Figure 3.1. The local density of states of the conduction band is then given by

$$d_c(E_{cl}) = 2 \cdot 2\pi (2m_e^*/h^2)^{3/2} \sqrt{E-E_{cl}}. \tag{3.33}$$

The density of valence band states assumes an analogous form. The energy E_{cl} is the local energy of the conduction band edge. The unperturbed local density of states for a donor level is given by

$$d_D(E_{Dl}) = N_D \delta(E-E_{Dl}), \tag{3.34}$$

where E_{Dl} denotes the local donor state energy level; and where N_D is the global average of the donor concentration. The densities of states are expressed per unit volume. The local energy levels are the

sums of the global averages and the random fluctuation, ΔE, e.g., for the conduction band edge

$$E_{cl} = E_{cav} + \Delta E. \tag{3.35}$$

As a consequence of this linear relationship the global density of states function is given as the convolution of the local density of states function with the probability density function of the random fluctuation, ΔE. Both Kane and Morgan found by different derivations that the random fluctuation is approximately normally distributed, and both found substantially the same variance. Kane proposes a standard deviation of the form

$$\sigma_o = [e^2/(4\sqrt{2}\pi\varepsilon)](4\pi\lambda\sum Z^2 N_Z)^{\frac{1}{2}}, \tag{3.36}$$

where N_Z is the density of impurities with charge Ze; and where λ is the Debye screening length. Morgan suggests a correction to (3.36) of the form

$$\sigma^2 = \sigma_o^2 \cdot 1.027 \exp-[2\cdot(11.3206\pi\lambda^3\sum Z^2 N_Z)^{-\frac{1}{2}}]. \tag{3.37}$$

Table 3.1 lists σ-values obtained from this equation.

Table 3.1. Energy Level Dispersion vs. Impurity Concentration

N_A+N_D =	10^{12}	10^{14}	10^{16}	10^{18}	10^{19}	10^{20}	10^{21}	$3\cdot10^{21}$	cm^{-3}
σ =	0,7	1,9	5,2	11,3	16,3	33,4	104	190	meV

Further derivations will be carried out for electrons in n-type silicon, but the results will also be applicable to holes and other semiconductors, with changes of the appropriate parameters.

As mentioned, the global density of states in the conduction band is given by the convolution

$$D_c(E) = 4\pi(\sqrt{m_e^*}/h)^3\sqrt{E-E_c}\,U(E-E_c)*(\sqrt{2\pi}\sigma)^{-1}\exp[-E^2/(2\sigma^2)], \qquad (3.38)$$

and the density of impurity states equals

$$D_D(E) = N_D\delta(E-E_D)*(\sqrt{2\pi}\sigma)^{-1}\exp[-E^2/(2\sigma^2)], \qquad (3.39)$$

where $U(x)$ denotes the unit step function; and where $\delta(x)$ is the Dirac delta function. At this point a few remarks concerning the screening length are in order. The screening length can be determined by introducing a test charge into a system of mobile carriers and performing a self-consistent field calculation. Let ϕ be the potential induced by the test charge. In a metal or semiconductor the density of carriers depends on the potential, ϕ, via the density of states function. Considering the test charge as a small perturbation, one can linearize the relationship between the carrier density and the potential and insert the result into Poisson's equation.

Let the concentration of electrons be $n[\phi(r)]$ and let $n(\phi=0) = n_o$ be the electron distribution in the charge neutral material. Neutrality implies, of course, that the electronic charge is exactly compensated by the bound positive charges of the ions; $n_o = p_o$. Now, for a small perturbation in the potential, the electron concentration, to first order, is

$$n(\vec{r}) \simeq n_o + \alpha\phi(\vec{r}); \quad \alpha = \partial n/\partial\phi_{\phi=0}. \tag{3.40}$$

Poisson's equation thus assumes the form

$$\nabla^2\phi(\vec{r}) = -(e^2/\varepsilon)[n(\vec{r})-n_o] = -(e^2/\varepsilon)\alpha\phi(\vec{r}). \tag{3.41}$$

For a point charge, q_o, at the origin, the solution to (3.41) is

$$\phi(\vec{r}) = [-eq_o/(4\pi\varepsilon r)]\exp-(r/\lambda); \quad \lambda = (\alpha e^2/\varepsilon)^{-\frac{1}{2}}. \tag{3.42}$$

In metals and extremely degenerate semiconductors with the Fermi level several kT above the conduction band edge, one obtains

$$n_M = (8/3)\pi h^{-3}(2m_e^*)^{3/2}(E_f-E_c-e\phi)^{3/2}, \tag{3.43}$$

$$\alpha_M = -(3/2)n_o/(E_f-E_c), \tag{3.44}$$

$$\lambda_M = [2\varepsilon(E_f-E_c)/(3n_o e^2)]^{\frac{1}{2}} = h(e^2 m_e^*/\varepsilon)^{-\frac{1}{2}}(3\pi^2 n_o)^{-1/6}. \tag{3.45}$$

If a semiconductor is doped with N_A acceptors per cm^3 and $N_D > N_A >> n_i$ donors per cm^3, the net density, $N_o = N_D - N_A$, of electrons will be distributed between the conduction band and the donor states. The Fermi level will be above the center of the band gap. Hence, the valence band and the acceptor states are practically filled and the few holes can be disregarded in comparison with N_o. When a test charge is introduced, the electrons in the donor states as well as those in the conduction band will redistribute and actively contribute to the screening. For a moderately doped semiconductor, where the Boltzmann approximation holds, one obtains

$$N_o = N_D - N_A \simeq n_i \exp[(E_f - E_{io})/(kT)], \tag{3.46}$$

$$N_s \simeq n_i \exp[(E_f - E_{io} - e\phi)/(kT)] \simeq (N_D - N_A)\exp[-e\phi/(kT)], \tag{3.47}$$

$$\alpha_s = -(N_D - N_A)/(kT), \tag{3.48}$$

$$\lambda_s = [\varepsilon kT/(e^2|N_D - N_A|)]^{\frac{1}{2}}. \tag{3.49}$$

The absolute difference of the impurity concentrations in the last equation has been introduced to reflect that screening by holes in p-material is equivalent to screening by electrons in n-material.

As the impurity concentration increases, the Fermi level approaches and crosses the impurity energy level. Then $N_s(\phi)$ varies less strongly with ϕ, and α_s will be smaller than indicated by (3.48). Hence λ_s as given by (3.49) is a lower bound for the screening length in semiconductors. For silicon at 300 K the screening lengths, λ_M and λ_s, are equal for $N_o = 4 \cdot 10^{19}$ cm^{-3}. For this electron concentration the Boltzmann approximation places the Fermi level at the conduction band edge, while Equation (3.43) places it 3/2 kT above the conduction band edge. Both calculations are inaccurate in this region. After some thought, one concludes that λ_M is also a lower bound on the screening length. Not knowing the true density of states function for a heavily doped semiconductor at this point, one can at best guess at the true value of the screening length. In analogy to many other situations where two asymptotes of a function cross, one would expect that the approximation

$$\lambda = (\lambda_M^2 + \lambda_s^2)^{\frac{1}{2}} \tag{3.50}$$

is closer to the true screening length than either λ_M or λ_s. Now all parameters of the density of states function are satisfactorily defined.

The next task is the calculation of the concentrations of electrons in the conduction band and in the donor states. To this end one has to integrate the density of states functions (3.38) and (3.39), weighted by the Fermi function,

$$f(E,E_f) = \{1+\exp[(E-E_f)/(kT)]\}^{-1}, \tag{3.51}$$

over all energies. The carrier density in the conduction band becomes

$$n_o = \int_E \int_{E'} D_c(E')(\sqrt{2\pi}\sigma)^{-1}\exp[-(E-E')^2/(2\sigma^2)]dE' \cdot$$

$$\cdot\{1+\exp[(E-E_f)/(kT)]\}^{-1}dE, \tag{3.52}$$

Similarly the concentration of neutral donors is

$$n_D = \int_E \int_{E'} D_D(E')(\sqrt{2\pi}\sigma)^{-1}\exp[-(E-E')^2/(2\sigma^2)]dE' \cdot$$

$$\cdot\{1+\exp[(E-E_f)/(kT)]\}^{-1}dE. \tag{3.53}$$

These integrals over physical entities have all the convergence attributes to allow the order of integration to be interchanged. One notes that the integral over E is the same in (3.52) and (3.53), when solved first. It reads

$$I = \int_{-\infty}^{\infty}(\sqrt{2\pi}\sigma)^{-1}\exp[-(E-E')^2/(2\sigma^2)]/\{1+\exp[(E-E_f)/(kT)]\}dE, \tag{3.54}$$

Infinite limits can be used because kT as well as σ are small compared with the band gap and with the width of the bands in all cases of interest here. Consider first the asymptotic behavior of the integral. For $\sigma \to 0$ the Gaussian degenerates to the Dirac delta function, and

$$\lim_{\sigma \to 0} I = \{1+\exp[(E'-E_f)/(kT)]\}^{-1} = f(E',E_f); \tag{3.55}$$

the Fermi function is reproduced. This is to be expected, because $\sigma = 0$ is the case with no band tailing.

When $kT \to 0$, $\sigma \neq 0$, the Fermi function degenerates to the unit step function, $U(E_f-E)$, and

$$\lim_{kT \to 0} I = \tfrac{1}{2}\mathrm{erfc}[(E_f-E')/(\sqrt{2}\sigma)]. \tag{3.56}$$

This function has overall features similar to the Fermi function. It is, however, not a faithful representation of the tails for moderate kT and σ values. When $E'-E_f >> \sigma, kT$, the Fermi function can be approximated by the Boltzmann factor, and the integral becomes

$$I = \int_{-\infty}^{\infty} (\sqrt{2\pi}\sigma)^{-1} \exp[-(E-E')^2/(2\sigma^2)] \exp[-(E-E_f)/(kT)]\,dE$$

$$= \exp-[(E'-E_f)/(kT) - \tfrac{1}{2}\sigma^2/(kT)^2]. \tag{3.57}$$

The Boltzmann function is reproduced with a shift in energy.

In the transition region, where no asymptotic approximations can be found by inspection, the integral (3.54) still in a way reproduces the Fermi function. As σ increases, the transition will become less steep. This behavior suggests approximation of the integral with a Fermi function, but with an effective temperature, $T_e(T,\sigma) > T$.

A study of the symmetries of the integrand of (3.54) reveals that $I(E'=E_f) = \frac{1}{2}$, independent of σ and kT. For extreme values of σ and kT one obtains a match of the first derivative at the point $E'=E_f$ for

$$\lim_{\sigma\to 0} kT_e = kT, \quad \text{and} \quad \lim_{kT\to 0} kT_e = \sigma\sqrt{\pi/8}. \tag{3.58}$$

The approximation $(kT_e)^2 = (kT)^2+\pi\sigma^2/8$ matches the derivative at the symmetry point very well. The approximation

$$(kT_e)^2 = (kT)^2+0,275\sigma^2, \tag{3.59}$$

however, provides a closer overall match of the effective Fermi function,

$$f_e = \{1+\exp[(E'-E_f)/(kT_e)]\}^{-1}, \tag{3.60}$$

with the integral I. The combination of (3,60) and (3,57) leads to the piece-wise approximation of the integral,

$$\tilde{I}(E'-E_f) = \begin{cases} 1-\exp[(E'-E_f)/(kT)-\frac{1}{2}\sigma^2/(kT)^2]; & E' < E_f-E_t; \\ \{1+\exp[(E'-E_f)/(kT_e)]\}^{-1}; & |E'-E_f| < E_t; \\ \exp[-(E'-E_f)/(kT)+\frac{1}{2}\sigma^2/(kT)^2]; & E' > E_f+E_t. \end{cases} \tag{3.61}$$

This function has maximum errors as summarized in Table 3,2, which also indicates the best choice of the transition point, E_t.

The range up to $\sigma/(kT_e) = 1,6$ covers most of the impurity concentrations and device operating temperatures of interest. Outside this region the errors increase quite rapidly and it becomes necessary to evaluate the integral (3,54) numerically. The values for $E_{tB}/(kT_e)$

Table 3.2. Error in Approximating Equation (3.54) by (3.61)

	$(\tilde{I}-I)/I$			
$\sigma/(kT_e)$	max%	min%	$E_t/(kT_e)$	$E_{tB}/(kT_e)$
0.2	0.59	-0.37	4.9	3.1
0.4	1.56	-1.42	4.3	3.2
0.6	2.47	-3.97	4.1	3.4
0.8	3.29	-4.77	4.0	3.6
1.0	4.55	-6.64	4.0	3.9
1.2	6.62	-8.41	4.0	4.3
1.4	8.90	-10.0	4.1	4.7
1.6	12.1	-11.4	4.2	5.1
1.8	16.6	-12.6	4.3	5.5
2.0	21.8	-13.7	4.5	5.9
2.5	41.3	-15.6	4.8	6.9
3.0	72.6	-16.8	5.2	8.0
3.5	119	-17.6	5.6	9.0
4.0	213	-18.2	6.0	10.0

indicate the point at which the Boltzmann approximation gives an error of 5%.

The remaining integrals in Equations (3.52) and (3.53) can now be written in terms of the approximation $\tilde{I}$. For very high impurity concentrations Equation (3.52), in view of (3.33), is

$$n_o = 2 \cdot 2\pi (\sqrt{2m_e^*}/h)^3 \int_{E_c}^{\infty} \sqrt{E-E_c}\,\tilde{I}(E-E_f)\,dE. \tag{3.62}$$

Blakemore [31, p. 360] approximates the integral by the expression $\Gamma(3/2)\exp[(E_f-E_c)/(kT)]/\{1 + 0.27\exp[(E_f-E_c)/(kT)]\}$, which is in error by only 3% at $E_f-E_c = 1.3kT$ and improves very rapidly as E_f declines. With this approximation the number of electrons in the conduction band becomes

$$\left.\begin{aligned} n_o &= N_{ce}\exp[(E_f-E_c)/(kTe)]/\{1+0.27\exp[(E_f-E_c)/(kTe)]\} \\ N_{ce} &= 2 \cdot 2\pi (\sqrt{2m_e^* kT_e}/h)^3 \Gamma(3/2). \end{aligned}\right\} \tag{3.63}$$

For lower concentrations, with the Fermi level about $4kT_e$ below E_c, the Boltzmann approximation becomes more accurate and the integral in (3.62) leads to

$$\left.\begin{aligned} n_o &= N_c\exp[(E_f-E_c)/(kT)+\tfrac{1}{2}\sigma^2/(kT)^2] \\ N_c &= 2 \cdot 2\pi (\sqrt{2m_e^* kT}/h)^3 \Gamma(3/2). \end{aligned}\right\} \tag{3.64}$$

The hole concentration is analogously given by

$$\left.\begin{aligned} p_o &= N_v \exp[(E_v-E_f)/(kT)+\tfrac{1}{2}\sigma^2/(kT)^2] \\ N_v &= 2\cdot 2\pi(\sqrt{2m_h^* kT}/h)^3\Gamma(3/2). \end{aligned}\right\} \tag{3.65}$$

The np-product thus becomes

$$n_o p_o = N_c N_v \exp-[E_g/(kT)-\sigma^2/(kT)^2] = n_i^2 \exp[\sigma^2/(kT)^2]. \tag{3.66}$$

As expected, the band-tailing causes a virtual narrowing of the band gap. The amount of shrinkage is simply $\sigma^2/(kT)$. The band shrinkage has the nature of a change of the semiconductor material and causes a modification of the intrinsic carrier density to

$$n_{ie} = n_i \exp[\tfrac{1}{2}\sigma^2/(kT)^2]. \tag{3.67}$$

In view of Equations (3.39), (3.53), and (3.61) the fraction of neutral donor atoms is

$$n_D/N_D = \begin{cases} 1-\exp[(E_D-E_f)/(kT)-\tfrac{1}{2}\sigma^2/(kT)^2]; & E_D < E_f-E_t; \\ [1+\exp[(E_D-E_f)/(kT_e)]^{-1}; & |E_D-E_f| < E_t; \\ \exp[-(E_D-E_f)/(kT)+\tfrac{1}{2}\sigma^2/(kT)^2]; & E_D > E_f+E_t; \end{cases} \tag{3.68}$$

n_D being the concentration of neutral donor atoms. If the Boltzmann approximation is also applicable to an impurity band, a shift of the impurity level by $\frac{1}{2}\sigma^2/(kT)$ toward the center of the band gap can account for the effects of impurity band broadening, i.e., the picture of band shrinkage due to high impurity concentration also holds for

the impurity levels; the average ionization energy of the impurity is unchanged.

When the concentration becomes so high that the Fermi level approaches the impurity level or the band edge to within 4.25 kT_e, Fermi-Dirac statistics must be used. In these cases the effects of the high concentrations are reflected in the effective temperature T_e. In summary, the effects of high concentration--band tailing and impurity band broadening--can be included in the classical Boltzmann approximation by simply shrinking the band gap by an appropriate amount. For very high concentrations, where the Boltzmann approximation is no longer applicable, the effects can be accounted for by the introduction of an effective temperature which is a simple function of the temperature and of the dispersion parameter, σ. In all these cases the analytical tractability is not impaired by the introduction of the corrections.

Previously Van Overstraeten, DeMan, and Mertens [32], Kleppinger [33], and possibly others have investigated the effects of high impurity concentrations on the carrier statistics, but they have taken recourse to numerical solutions. The derivations of this section culminate in simple and quite accurate analytical approximations. In this treatment the distinction between conduction band electrons and electrons bound to the donor atoms is maintained, although the energies of the states in the tail of the conduction band and of the states of the impurity band overlap. This distinction is consistent with the origin of the dispersion of the energy levels as described at the beginning of the section. Even though their energies are equal, an electron localized at

a donor atom will still not contribute to conduction, while the conduction band electron will.

The separation of the mobile and bound particles requires a consistent specification of the mobility and diffusivity. A review of the literature reveals for example, that the mobilities as specified by Grove [8, p. 110] are based on the assumption that even at the highest concentrations the impurity atoms are totally ionized, i.e., that all electrons participate in the conduction processes. Sze [25, p. 40], on the other hand, specifies the mobility on the basis of only those particles which contribute to the conduction. His curves account for the considerable "freeze-out" which occurs at high concentrations, but they still do not reflect band tailing, and require further modifications at the high concentration end. Since all mobility curves in the literature are results of experiments, one must in principle find out from the original publications how the experimental results have been interpreted, and on what assumptions about the free carrier concentration the mobility calculations are based.

3.4 Equilibrium Carrier Distributions in Nonuniformly Doped Material

The first problem in modeling the electrical behavior of semiconductor regions with impurity profiles is the determination of the carrier distributions at thermostatic equilibrium. From them one can derive the electrical and the chemical potentials at equilibrium, which, in turn, serve as the starting parameters for the calculation of non-equilibrium states by the methods suggested in Section 3.2. Furthermore, the majority carrier distribution at thermostatic equilibrium is

frequently representative of the nonequilibrium distribution because practicable deviations from equilibrium cause only small perturbations of the equilibrium profiles, except in the neighborhood of metallurgical junctions.

The equilibrium carrier distributions satisfy the equations of the carrier statistics and Poisson's equation. The analytically simplest case of a moderately doped semiconductor, where the impurities can be considered fully ionized, and where Boltzmann statistics apply, yields good insight into the essential features of the multi-dimensional carrier distributions. Poisson's equation for such cases is

$$\nabla^2\phi_o = -\rho/\varepsilon = (e/\varepsilon)[n_o(\vec{r})-p_o(\vec{r})-N(\vec{r})], \tag{3.69}$$

where ϕ_o is the electrostatic potential in thermostatic equilibrium; where n_o and p_o are the equilibrium electron and hole concentrations; and where N is the excess of the donor concentration over that of the acceptors, $N = N_D-N_A$. The following derivations will again assume n-regions, with the understanding that analogous relations hold for p-regions.

In the Boltzmann approximation the carrier concentrations can be written as

$$n_o(\vec{r}) = n_i\exp[e\phi_o(\vec{r})/(kT)], \tag{3.70}$$

$$p_o(\vec{r}) = n_i\exp[-e\phi_o(\vec{r})/(kT)], \tag{3.71}$$

which is consistent with Equations (3.19) and (3.20). Now consider a sample of homogeneously doped material with a net impurity concentration just equal to $N(\vec{r}_1)$. Such a sample has a constant potential ϕ_{oh}

throughout, and carrier concentrations are

$$n_{oh} = n_i \exp[e\phi_{oh}/(kT)], \tag{3.72}$$

$$p_{oh} = n_i \exp[-e\phi_{oh}/(kT)], \tag{3.73}$$

h indicating parameters of the homogeneous sample. Since the homogeneous sample is charge neutral, its carrier concentrations satisfy the equation

$$n_{oh}-p_{oh}-N(\vec{r}_1) = 0, \tag{3.74}$$

where the argument $\vec{r}_1$ is simply a reflection of the way in which the sample has been chosen, and is not a running variable in the homogeneous sample. Consider now calculating the concentrations n_{oh} and p_{oh} for all concentrations $N(\vec{r})$ which occur in the nonuniform profile, where $\vec{r}$ is simply a bookkeeping index, for the moment. Since Equations (3.72) to (3.74) hold point by point, one can write

$$n_{oh}(\vec{r})-p_{oh}(\vec{r})-N(\vec{r}) = 0, \tag{3.75}$$

In view of their construction, let $n_{oh}(\vec{r})$ and $p_{oh}(\vec{r})$ be called the virtual carrier profiles, in contrast to the real profiles $n_o(\vec{r})$ and $p_o(\vec{r})$. Assuming that the virtual profiles and the real profiles differ only insignificantly from each other, one may write

$$n_o(\vec{r}) = n_{oh}(\vec{r})+\nu(\vec{r}), \tag{3.76}$$

$$p_o(\vec{r}) = n_i^2/n_o(\vec{r}) \simeq p_{oh}(\vec{r})-\nu(\vec{r})p_{oh}(\vec{r})/n_{oh}(\vec{r}), \tag{3.77}$$

In view of Equation (3.75), Poisson's equation can thus be written in the form

$$\nabla^2\phi_o(\vec{r}) = (e/\varepsilon)\nu(\vec{r})[1-p_{oh}(\vec{r})/n_{oh}(\vec{r})] \simeq e\nu(\vec{r})/\varepsilon. \tag{3.78}$$

On the other hand one suspects that for strongly extrinsic material $\phi_o(r)$ differs very little from $\phi_{oh}(r)$. If this is true, one may write

$$\phi_o(\vec{r}) \simeq \phi_{oh}(\vec{r}) = (kT/e)\ln[n_{oh}(\vec{r})/n_i], \tag{3.79}$$

$$\nabla\phi_o = -\vec{E} \simeq (kT/e)(\nabla n_{oh})/n_{oh}, \tag{3.80}$$

$$\nabla^2\phi_o \simeq (kT/e)[(\nabla^2 n_{oh})/n_{oh}-(\nabla n_{oh})^2 n_{oh}^2]. \tag{3.81}$$

The combination of (3.78) with (3.81) results in

$$\nu/n_{oh} \simeq -[\varepsilon kT/(e^2 n_{oh})][(\nabla^2 n_{oh})/n_{oh}-(\nabla n_{oh})^2/n_{oh}^2], \tag{3.82}$$

This equation expresses the relative deviation of the true majority carrier profile from the virtual one. The first factor on the right is the square of the Debye screening length, λ_{oh}, of the virtual majority carrier concentration. The second factor is a function of the virtual majority carrier concentration, its gradient and its Laplacian. It depends largely on the features of the impurity profile. In strongly extrinsic material (in silicon for N between about 10^{14} cm^{-3} and 10^{18} cm^{-3}), $n_{oh} \simeq N$, and one can replace n_{oh} by N in Equation (3.82) with little error. From the second chapter of this work one concludes that, except near metallurgical junctions, the impurity profiles generated by diffusion processes are germane to the one-dimensional Gaussian or the complementary error function profiles. It is therefore illustrative to determine ν/n_{oh} for these two profiles. Using the inequality [15, eqn. 7.1.13], $u+(u^2+2)^{\frac{1}{2}} > (2/\sqrt{\pi})\exp(-u^2)/\mathrm{erfc}(u) > u+(u^2+4/\sqrt{\pi})^{\frac{1}{2}}$, and some straightforward algebra, one obtains for

$$\text{Gaussian:}\quad \nu/n_{oh} = -2(\lambda_{oh}/\lambda_D)^2, \tag{3.83}$$

$$\text{Erfc:}\quad \nu/n_{oh} \simeq -(4/\pi)(\lambda_{oh}/\lambda_D)^2, \tag{3.84}$$

where $\lambda_D = (4Dt)^{\frac{1}{2}}$ is the diffusion length of the impurity diffusion process and is independent of the position, while λ_{oh} is position dependent. As an example, in a steep predeposition diffusion profile, as might be used in a dipped emitter process, λ_D might be as small as 200 nm. The Debye length as a function of the impurity concentration, derived from Equation (3.50), is given in Table 3.3. Therefore, in the above steep profile ν/n_{oh} is less than 0.01 in the regions where N is larger than $3 \cdot 10^{16}$ cm^{-3}. For flatter profiles the region of less than 1% deviation will extend to even lower concentrations.

It is worth noting that the assumptions made in the course of the derivation, i.e., Boltzmann approximation, strongly extrinsic material, and small deviation of n_{oh} from n_o, and of ϕ_{oh} from ϕ_o, are fully met in the range of concentrations up to about 10^{18} cm^{-3}. At higher concentrations n_{oh} will be smaller than N due to freeze-out, and the carrier profile will be flatter than the impurity profile. This can be accounted for by a reduced, effective Debye length and a reduced, effective diffusion length; two corrections which tend to counteract. The deviation of the true profile from the virtual one will not be significantly affected and it is much smaller than 1% to begin with, due to the exceedingly small Debye length.

Table 3.3. Debye Screening Lengths

N =	10^{12}	10^{14}	10^{16}	10^{18}	10^{20}	10^{22}	cm^{-3}
λ_{oh} =	4140	414	41.5	4.32	0.70	0.27	nm

Thus, the derivations of this and the previous section have provided the tools for calculating the equilibrium carrier profiles in all regions of a semiconductor device structure, except in the transition regions near metallurgical junctions. Equations (3.82), (3.83), and (3.84) illustrate that the net uncompensated charges are tiny fractions of the free charges represented by the majority carrier profiles. The material is practically charge neutral, even in regions of steep impurity profiles.

Near the junctions, and also near contacts and semiconductor surfaces, the derivation of this section breaks down. Near junctions the material is no longer strongly extrinsic, while the contacts and the semiconductor surfaces represent discontinuous changes of the profiles.

3.5 Transition Regions of pn-Junctions in Thermostatic Equilibrium

In the planar semiconductor technology metallurgical junctions are produced by diffusing a profile with a higher surface concentration, a steeper gradient, and opposite polarity through a diffusion window into an already existing profile in the substrate material. The pn-junction then forms a three-dimensional surface where the donor and acceptor concentrations are equal. In the transition region of the junction the net impurity concentration, $N = |N_D - N_A|$, assumes low values and, except for the Boltzmann statistics, the approximations of the previous section are no longer valid. One is then faced with solving the partial differential equation

$$\nabla^2 \phi_o(\vec{r}) = (e/\varepsilon)\{2n_i \sinh[e\phi_o/(kT)] - N(\vec{r})\}, \qquad (3.85)$$

This problem has no tractable analytic solutions for curved junctions of arbitrary profiles. Short of numerical solutions one will have to make judicious approximations in order to make the problem tractable. To gain better insight into the nature of the solutions, it is again helpful to analyze planar junctions formed by specialized impurity profiles. The one-dimensional profiles from diffusions are awkward to handle, but an inspection of plots of the logarithm of the Gaussian and of the complementary error function shows that exponential functions are good local approximations of these profiles. Therefore, consider a junction formed by exponential donor and acceptor profiles of the form

$$N_D = N_j \exp(x/L_D); \quad N_A = N_j \exp(x/L_A), \tag{3.86}$$

where the junction is the origin of the coordinate system; and where positive and negative values of the parameters L_A and L_D indicate concentrations which increase or decrease with increasing x.

It is instructive to consider a case for which Equation (3.85) has an exact solution. For a symmetrical, planar junction with $L_D = -L_A = L > 0$, the equation becomes

$$d^2\phi_o/dx^2 = (e/\varepsilon)\{2n_i \sinh[e\phi_o/(kT)] - 2N_j \sinh(x/L)\}, \tag{3.87}$$

which has a particular solution

$$N_j = n_i; \quad \phi_o = kTx/(eL). \tag{3.88}$$

Hence, for symmetrical exponential profiles with the concentrations at the metallurgical junction equal to n_i, the transition region is

exactly charge neutral. Such profiles never occur in practice, because the lowest achievable impurity concentrations are a few orders of magnitude higher than n_i. In these cases a transition region forms which is depleted of majority carriers. Far enough away from the junction one or the other profile becomes dominant. On the n-side the Equation (3.87) eventually simplifies to

$$\partial^2\phi_o/dx^2 = (e/\varepsilon)\{n_i\exp[e\phi_o/(kT)]-N_j\exp(x/L)\}, \tag{3.89}$$

with solution

$$\left.\begin{aligned} &\phi_n(x) = \phi_{no}-E_n x; \\ &\phi_{no} = (kT/e)\ln(N_j/n_i); \quad E_n = -kT/(eL). \end{aligned}\right\} \tag{3.90}$$

Similarly, the asymptotic potential on the p-side is

$$\phi_p(x) = -\phi_{no}-E_n x. \tag{3.91}$$

Figure 3.2 illustrates the two asymptotic potentials and the potential in the transition region. One notes that the potential, being the second integral of the space charge, must be smooth everywhere. One observes also that $\phi_o(x)$ must be concave ($d^2\phi_o/d_x{}^2 > 0$) on the p-side and convex ($d^2\phi_o/dx^2 < 0$) on the n-side. Since $\rho = -(\varepsilon/e)d^2\phi_o dx^2$), a negative space charge is required on the p-side, and a positive one on the n-side. The ionized impurities thus must dominate in the transition region; the transition region is a depletion region. The single condition which makes it a depletion region is that $N_j > n_i$.

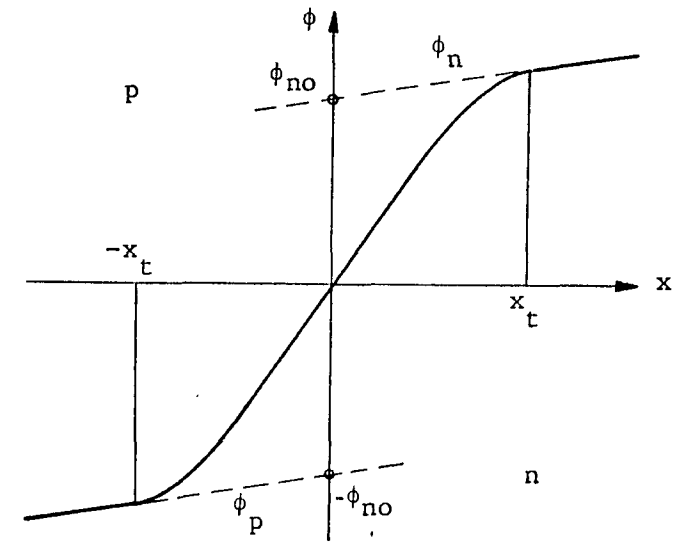

Figure 3.2. Asymptotic Potentials and Transition Potential in a PN Junction with Symmetrical Exponential Profiles

The question of the degree of depletion and of the detailed shape of $\phi_o(x)$ arises next. The classical approximation is to neglect free carriers entirely. Dropping the term $2n_i\sinh[e\phi_o/(kT)]$ from Equation (3.87), and integrating the equation twice, gives the potential in the region $-x_t < x < x_t$ as

$$\phi_o(x) = 2(e/\varepsilon)N_j L^2[(x/L)\cosh(x_t/L)-\sinh(x/L)]+kTx/(eL), \qquad (3.92)$$

This solution satisfies the boundary condition on the electric field at x_t. The parameter, x_t, can be calculated by also invoking the

boundary condition on the potential at x_t. Expanding the hyperbolic functions into a Taylor series about the origin, in the hope of obtaining a simplifying approximation, leads to

$$\phi_o(x_t) = (2eN_jL^2/\varepsilon)[x_t^3/(3L^3)+x_t^5/(30L^5)+x_t^7/(840L^7)+...]$$

$$= (kT/e)\ln(N_j/n_i). \tag{3.93}$$

If $x_t/L < 1$, the truncation of the series after the first term introduces less than 10% error. This truncation is tantamount to approximating the exponential impurity profiles by linearly graded profiles. With this approximation one obtains

$$x_t = [3/2)LL_j\ln(N_j/n_i)]^{1/3}, \tag{3.94}$$

where L_j is the Debye screening length associated with the concentration N_j. The condition $x_t < L$ is, for example, satisfied for concentrations $N_j = 10^{14}$ cm^{-3}, 10^{16} cm^{-3}, and 10^{18} cm^{-3}, when $L > 1500$ nm, 185 nm, and 22 nm, respectively.

How good is the depletion approximation? Since $\phi_o(x_t) = \phi_n(x_t)$, in reality one ought to have

$$n_i\sinh[e\phi_o(x_t)/kT)] = N_j\sinh(x_t/L), \tag{3.95}$$

Because $N_j >> n_i$, the argument of the hyperbolic sine on the left must be much larger than the corresponding argument on the right, which is of the order of unity. Therefore, while the net impurity concentration rises almost linearly between 0 and x_t, the majority carrier concentration starts much flatter at the junction and later must curve upwards

very strongly in order to approach the impurity concentration at x_t. The majority carrier concentration is thus insignificantly low in comparison to the impurity concentration over most of the transition region.

The depletion approximation leads to a lower bound estimate of the width of the transition region. In reality the space charge does not drop to zero abruptly at x_t, but starts tapering off about two Debye lengths before x_t, and charge neutrality is reached about two Debye lengths beyond x_t. The relevant concentration for calculating the Debye length in this case is $|N(x_t)|$.

The planar technology rarely produces symmetrical junctions; in fact, in the majority of cases the donor and acceptor concentrations slope in the same direction. Figure 3.3 illustrates the two asymptotic potentials and the potential through the transition region of such a junction, with the profiles given by Equations (3.86) with $L_D > L_A > 0$. In contrast to the symmetrical junction, the electric fields at x_n and x_p are of opposite direction, both pointing toward the junction. The junction therefore contains a net negative charge per unit area,

$$q = \varepsilon[E(x_n)-E(x_p)]. \tag{3.96}$$

Since $N(x) > -N(-x)$, the potential is curved more strongly at each point $x > 0$ than at the corresponding point $-x$. It can only match up smoothly with both asymptotes if it is positive at $x = 0$. Nevertheless, since also in these junctions $N_j >> n_i$, the depletion approximation is still good. The calculation of the space charge widths, x_n and x_p, is more involved, however. It requires numerical solutions.

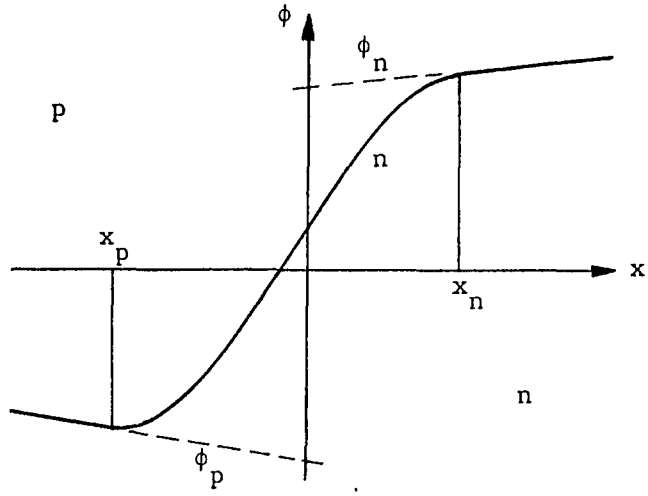

Figure 3.3. Asymptotic Potentials and Transition Potential in a PN Junction with Asymmetric Exponential Profiles

The depletion approximation reduces the calculation of the potential, ϕ, in the transition region to a problem of electrostatics. It is not a boundary value problem, though, because the boundaries, i.e., the curved interfaces between the space charge region and the charge neutral regions, are not known a priori. General solutions of this type of problem are not known. However, substantially curved junctions usually occur near the semiconductor surface, where the impurity concentrations and the gradient of N are high. This makes the width of the depletion region small in comparison to the radius of

curvature. Therefore one can locally treat a curved junction with the techniques for planar junctions with little error.

3.6 Interfaces of the Semiconductor with Metals or Insulators

All semiconductor surfaces represent interfaces, either with insulators, metals, or with other semiconductors.

In the semiconductor device technology one finds predominantly two types of interfaces. Most of the semiconductor surface is covered with high quality solid insulators. In the silicon technology amorphous, thermally grown silicon dioxide is used almost exclusively as the first layer of insulator. The areas not covered by insulators are contact areas and areas of Schottky barrier diodes. Aluminum is by far the most frequently used metal for ohmic and Schottky barrier contacts. Polycrystalline and heavily doped silicon is sometimes used in place of metal interconnections in the silicon IGFET (insulated gate field effect transistor) technology.

The difference between an ohmic contact and a Schottky barrier contact is phenomenological. An ohmic contact, as the name might suggest, exhibits--within measurement accuracy--proportionality between voltage drop and current, for both polarities. In contrast, the Schottky barrier contact exhibits a nonlinear relationship between current and voltage of the form

$$I_c = I_o\{\exp[eV_c/(kT)]-1\}. \tag{3.97}$$

The insulators and contact materials link the semiconductor material to the outside world, and it becomes necessary to define the potentials in a less parochial way than was convenient so far.

The energy which has to be imparted to an electron in order to emit it from the conduction band edge in the semiconductor to a surrounding vacuum is the electron affinity, χ. It is the binding energy of the electron to the semiconductor crystal and is independent of the impurity concentration as long as no band tailing has to be considered. This assertion might seem inconsistent in the case of a pn-junction because the conduction band edge on the p-side is at a higher potential than on the n-side. However, the electrostatic potential difference between the p- and the n-sides extends into the vacuum. Thus, an electron emitted barely beyond the surface of the n-region is still at a lower potential energy than an electron near the surface of the p-region. The electrostatic potential difference and the difference of the potential energies of electrons have opposite sign because of the negative charge of the electrons.

In highly doped material, where band tailing becomes significant, the electron affinity is increased by one half of the band shrinkage defined in Section 3.3. To see this consider that the vacuum energy level is everywhere equidistant from the intrinsic energy level. With increasing impurity concentration the virtual conduction band edge, $E_c - \frac{1}{2}\sigma^2/(kT)$, sinks toward the intrinsic level by the amount $\frac{1}{2}\sigma^2/(kT)$. The distance from the vacuum level to the virtual conduction band edge, χ, thus increases by the same amount.

The energy which must be furnished to an electron in order to lift it from the Fermi level of a metal to the vacuum level is the work function, W.

The Fermi level of a metal is offset by the difference $W - \chi$ from the conduction band edge of a disjoint semiconductor. For example, the work function of aluminum is given as 4.25 eV and the electron affinity of silicon is given as 4.05 eV [25, p. 366]. Figure 3.4 illustrates the band structures of isolated and of joined blocks of aluminum and silicon.

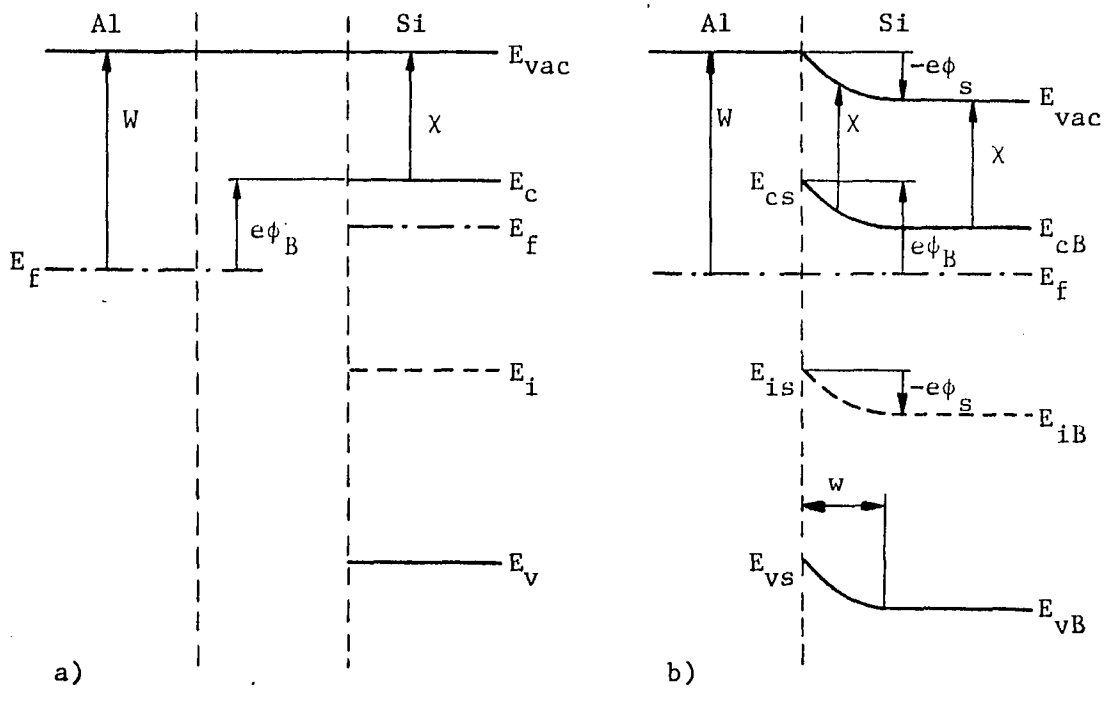

Figure 3.4. The Band Structure of Aluminum and Silicon; (a) When They are Isolated; (b) When They are in Contact

The barrier potential,

$$\phi_B = (W-\chi)/e, \tag{3.98}$$

remains unchanged upon joining. Therefore, in order to obtain a constant Fermi level as required for thermostatic equilibrium, a surface potential, ϕ_s, with respect to the interior of the semiconductor must build up;

$$\phi_s = (E_{is}-E_{iB})/e = [E_{iB}/e-\phi_B-(kT/e)\ln(N/n_i)]. \tag{3.99}$$

The definitions of the symbols are implied by Figure 3.4. The surface potential is generated by a space charge layer of width, w, in the semiconductor. A surface charge of opposite polarity, equal in magnitude to the space charge layer is induced on the metal surface in order to keep the electric fields in the metal as well as in the charge neutral interior of the semiconductor at zero. Under the assumed idealized interface conditions ϕ_s vanishes for $N \simeq 2 \cdot 10^{16}$ cm^{-3}; it is negative for larger N and positive for smaller N. A negative surface potential causes the bands to bend up, as shown in the figure. In this case the conduction band edge forms a barrier for the electrons. Similarly, a positive surface potential, obtained with $N < 2 \cdot 10^{16}$ cm^{-3}, bends the bands down introducing a barrier for holes. The width of the barrier can be determined with a formulation given, e.g., by Sze [25, pp. 429 ff]. The results are listed in Table 3.4.

A barrier for the majority carriers leads to a diode characteristic, unless it is thin enough that the carriers can tunnel through it. One can estimate the relative importance of tunnelling by

Table 3.4. Schottky Barrier Width as a Function of $N = N_D - N_A$

$N =$	10^{20}	10^{19}	10^{18}	10^{17}	-10^{16}	-10^{17}	-10^{18}	-10^{19}	-10^{20}	cm^{-3}
$w =$	1.6	4.1	10	19	88	78	31	11	3.4	nm

comparing the tunnelling probability of an electron with the probability that it is thermally excited to an energy $e\phi_B$ above the conduction band edge. The barrier can approximately be described by

$$E_c(x) - E_{cB} = (e^2 N/\varepsilon)(x-w)^2. \tag{3.100}$$

The tunnelling probability, in the Wentzel-Kramer-Brillouin formulation, is given by

$$P_T = \exp -\left\{2\int_0^w [2m_e^*(2\pi/h)^2[E_c(x) - E_{cB}]]^{\frac{1}{2}}dx\right\}$$

$$= \exp -[2w^2e^2m_e^*(N/\varepsilon)(2\pi/h)^2]^{\frac{1}{2}}, \tag{3.101}$$

while the probability of an electron jumping the barrier is given by the Boltzmann factor

$$P_j = \exp -[e\phi_B/(kT)]. \tag{3.102}$$

Some ratios, P_T/P_j, as a function of N are recorded in Table 3.5. The ratio P_T/P_j falls extremely rapidly with the magnitude of the net impurity concentration. When it is larger than unity the junction is ohmic, otherwise it forms a Schottky barrier diode. Table 3.5 suggests the criterion for ohmic contacts; impurity concentrations of

Table 3.5. Relative Tunnelling Probabilities as a Function of N

$N =$	10^{20}	$3 \cdot 10^{19}$	10^{19}	10^{18}	10^{17}	-10^{18}	-10^{19}	-10^{20}	cm^{-3}
$P_T/P_j =$	26	0.65	$7 \cdot 10^{-3}$	10^{-7}	$7 \cdot 10^{-11}$	$3 \cdot 10^{-60}$	$1.5 \cdot 10^{4}$	$6 \cdot 10^{8}$	

the order of 10^{20} cm^{-3}. One can make Schottky barrier diodes with aluminum to n- or to p-material interfaces when the impurity concentrations are below about 10^{19} cm^{-3} and 10^{18} cm^{-3} respectively, but since the barrier for n-material is much lower, it is much more difficult to obtain n-type Schottky barrier diodes.

In practice the aluminum-silicon interface is not nearly as uncomplicated as assumed for this discussion. Interface layers of insulators, predominantly of silicon dioxide-alumina complexes are often present. One also expects the bandgap to decay gradually toward the interface due to an increasing density of allowed states in the band gap caused by the increasing aluminum concentration. During customary annealing cycles of typically 10 to 20 minutes at 500°C, the alloying of the aluminum and silicon, and some diffusion are furthered. The literature is full of discussions of interface phenomena. The details of the interface structures depend very much on the processes used, and much of the theories is still in an unsettled state. It is clear, though, that for ohmic contacts the interface layers must be transparent to the carriers, and that in thermostatic equilibrium the net effect of

charges in the interface layer will be that of a dipole layer which changes the effective barrier height.

Much work has in recent years been devoted to understanding and controlling the phenomena which occur at interfaces of the semiconductors with insulators, because they are crucial to the characteristics of IGFETs. The properties of these interfaces depend very strongly on the fabrication processes. They can in principle be described in terms of an effective surface charge layer and a density of surface states. The net effect at thermostatic equilibrium is a surface charge density of the order of 10^{11} electron charges per cm^2 for well passivated surfaces. The surface charge induces a space charge layer in the semiconductor with which it forms a dipole layer. This dipole layer, in turn, produces a potential difference between the surface and the interior of the semiconductor. The potential difference depends on the surface charge density and on the impurity profile and it usually varies over the surfaces. Experience indicates that the surface phenomena are much less critical for bipolar devices than for FET devices.

From the theory of thermodynamics one concludes that at thermostatic equilibrium the net generation rate of electron-hole pairs is everywhere zero. This does not mean that no pairs are generated, but only that the generation and recombination rates are exactly equal. Conceptually, the exact balancing between the up and down transitions of electrons between the valence band and the conduction band is not an accidental consequence of the transition probabilities, but rather, it and the Fermi statistics are constraints

which govern the relationship between these probabilities. The transition rates themselves are strongly dependent on microscopic material properties; they increase with the densities of defects or impurities which create localized allowed states with energies near E_i. Such states are usually more concentrated near the semiconductor surface. At metal contacts their concentration is extremely high. On the metal side it is given roughly by the density of states near the Fermi level. A consequence of these conditions are the life times of carriers. The life time of minority carriers in silicon is of the order of microseconds, while the life time of an electron in a state near the Fermi level in a metal is of the order of picoseconds.

While differences in the transition rates have no explicit effects on the equilibrium concentrations of carriers, they do so under nonequilibrium conditions. Locations of high transition rates tend to restore nonequilibrium concentrations more strongly toward equilibrium. In particular, at contacts the electrochemical potentials for holes and electrons are forced extremely closely to the equilibrium condition, so that the boundary condition is always $np = n_i^2$. At interfaces of the semiconductor with insulators the transition rate is usually higher than in the interior. The measure of the increased rate is the surface recombination velocity. It expresses the number of minority carriers which recombine per cm^2 and per second for a unit of excess of minority carrier concentration.

Under nonequilibrium conditions the surface recombination velocity causes carrier fluxes toward the surface which are proportional to the deviation of the np product at the surface from n_i^2.

In conclusion of this section it might be noted that the basic phenomena which govern the transport of carriers through the semiconductor devices have been reviewed and outlined. The boundary conditions at the surfaces of the semiconductor region which can be influenced from the outside in order to operate the device have been discussed, so that the carrier transport problems are now adequately and consistently defined. The nonlinearities in the transport equations unfortunately prevent the solution of these problems by analytic techniques with sufficient accuracy. The next chapter investigates the variational principles of transport theory, which can be used for modeling with the aid of the finite element method. The finite element method appears to be the only technique which may lead to adequate solutions of three-dimensional problems.

CHAPTER 4

VARIATIONAL PRINCIPLES OF THE DISTRIBUTIONS AND FLUXES OF CARRIERS

4.1 Outline of the Chapter

In the preceding chapter the transport phenomena were formulated in terms of deviations from the state of thermostatic equilibrium. This motivated the investigation of approximate analytic methods for calculating equilibrium distributions of carriers. These techniques proved applicable to strongly extrinsic semiconductor regions, but they fail to describe the carrier profiles in two- and three-dimensional junction regions. For these regions one must resort to numerical calculations.

Finite difference schemes have traditionally been used for performing such calculations for one- and two-dimensional structures. The strong nonlinearities in the partial differential equations almost always confronted the investigators with difficulties with the numerical stability of the computation algorithms. No truly three-dimensional calculations have been carried out because of the excessive number of nodes which would be necessary to achieve numerical stability in the finite difference calculations.

The finite element method promises feasibility of such calculations because it can achieve numerical stability with a coarser and irregular grid of nodes. The finite element method has only recently found scant interest for application to carrier distribution and carrier

transport problems, e.g., by Buturla and Cottrell [34] and Cottrell and Buturla [35]. Buturla and Cottrell use an iterative scheme consisting of a finite element solution of Poisson's equation for the charge distribution and a finite difference solution of the flux distribution. The iteration is stopped when self-consistency is reached. The method is applied to two-dimensional FET structures.

The variational formulations for the finite element schemes are usually derived by mathematical considerations and comparisons with analogous or similar problems. Since such an approach is not without pitfalls, as already demonstrated in the second chapter, the following sections will be devoted to the derivation of the variational principles from the laws of physics.

In the next section a variational description of the equilibrium distributions of carriers and of the electrostatic potential in an arbitrarily doped semiconductor region will be derived. The development will show that the variational formulation which corresponds to Poisson's equation in electrostatics is the statement that the state of minimum Helmholtz free energy is the thermostatic equilibrium state.

The third section considers the variational formulation of steady state carrier currents. It is demonstrated that the variation of the entropy generation rate indeed leads to the well-known particle current equations, provided that the deviations from thermostatic equilibrium are sufficiently small. In an isothermal region the concept of minimum entropy generation rate is fully equivalent to minimum power dissipation. Thus the latter condition, proved already by Maxwell [36, pp. 407-408] for linear resistive circuits and by Millar [37] for

nonlinear resistive circuits, is also obtained from the postulates of thermodynamics. An inspection of the power dissipation equation in the light of the discussions of the second section clearly shows the stringent limits to be imposed on the magnitude of the deviations from equilibrium.

4.2 The Variational Formulation of the Thermostatic Equilibrium State of a Semiconductor

A piece of nonuniformly doped semiconductor material in thermostatic equilibrium gives rise to an interesting boundary value problem in electrostatics. It is a region which contains a distribution of immobile charge, formed by the ionized impurity atoms. It also contains two distributions of mobile charges of opposite polarities, the holes and the electrons. Although there is attraction between these charges, there is no net recombination; on a macroscopic scale the charge distributions are totally stationary. The goal of this section is to find a variational principle which describes the three charge distributions, and which can be exploited to calculate them numerically by means of the finite element method.

Deriving the variational principle by mathematical considerations failed in this case. Since only the careful reflection on the physical laws governing the system of charges led to success, it is felt worthwhile to retrace some of the developments.

Consider first a simpler boundary value problem in electrostatics; the determination of the potential distribution in an insulator with dielectric permittivity $\varepsilon(\vec{r})$, due to a charge distribution $\rho(\vec{r})$.

The boundary of the space may be at infinity or on conducting electrodes, with one of them completely enclosing the insulating region. The boundary may also just be an arbitrary closed surface on which suitable boundary conditions are specified.

The simplest boundary conditions are those where the potential is specified everywhere on the bounding surface. In this case the system is closed to an exchange of electrostatic energy with its surrounding, while it can still exchange heat. It is well known that under these conditions the Helmholtz free energy is a minimum when the system is in thermostatic equilibrium. This energy can be expressed in two forms,

$$W = \tfrac{1}{2}\iiint_V \rho(\vec{r})\phi(\vec{r})d\vec{r} = \tfrac{1}{2}\iiint_V \varepsilon(\vec{r})(\nabla\phi)^2 d\vec{r}, \tag{4.1}$$

where V is the volume of the insulating region, exclusive of the surfaces of conducting boundaries; and where ϕ is the electrostatic potential. To make the free energy a function of ρ, ϕ, and $\nabla\phi$, one writes

$$W = \iiint_V [\rho(\vec{r})\phi(\vec{r}) - \tfrac{1}{2}\varepsilon(\vec{r})(\nabla\phi)^2]d\vec{r} = \iiint_V w d\vec{r}. \tag{4.2}$$

Now it is necessary to show that the function $\phi(\vec{r})$ which minimizes W also satisfies Poisson's equation. One also needs a formulism for treating more complicated boundary conditions.

Feynman, Leighton, and Sands [38, Chapter 19] wrote one of the most lucid descriptions of the variational procedures for minimizing functionals like (4.2). The essence can again be formulated in terms

of the Euler-Lagrange equation

$$[\partial/\partial\phi - \nabla\cdot\partial/\partial(\nabla\phi)]w = 0. \tag{4.3}$$

As has already been indicated in the second chapter, the partial derivatives with respect to the functions ϕ and $\nabla\phi$ must in this case be calculated as if ϕ and $\nabla\phi$ were independent variables, and the variable $\nabla\phi$ has further to be treated as if it were a scalar. Executing the differentiations results in

$$\rho + \nabla\cdot(\varepsilon\nabla\phi) = 0, \tag{4.4}$$

which is indeed Poisson's equation.

It is worth pointing out that the function $\phi(\vec{r})$ which minimizes W also minimizes the free energy in an arbitrary subregion of the volume, V.

In deriving the Euler-Lagrange equation one finds another condition which has to be satisfied in those regions of the boundary surface where ϕ_s is not specified. This condition is

$$\nabla\phi\cdot d\vec{s} = 0. \tag{4.5}$$

It is thus advantageous to have those regions of the boundary, on which ϕ_s is not specified, spanned by flow lines of the potential.

More complex boundary conditions will make the term $\nabla\phi\cdot ds$ nonzero. Such conditions usually must be introduced with the aid of Lagrange multipliers. These will add terms to Equation (4.5) which just compensate for the nonzero term $\nabla\phi\cdot d\vec{s}$.

The initial success of the variational formulation might suggest to solve Equations (3.15), (3.16), and (3.17) for n and p and

to describe the charge density in a semiconductor by

$$\rho(\phi) = e[p(\phi)-n(\phi)+N(\vec{r})]$$

$$= e\{\exp[-e\phi/(kT)]-\exp[e\phi/(kT)]+N(\vec{r})\}, \tag{4.6}$$

where $N(\vec{r}) = N_D^+(\vec{r})-N_A^-(\vec{r})$ is the net concentration of positive impurity ions. The resulting energy density would then be

$$w = \phi(\vec{r})\rho(\phi)-\tfrac{1}{2}\varepsilon(\vec{r})(\nabla\phi)^2 \tag{4.7}$$

and would lead to an Euler-Lagrange equation

$$\phi\partial\rho/\partial\phi+\rho(\phi)+\nabla\cdot(\varepsilon\nabla\phi) = 0, \tag{4.8}$$

which is clearly not Poisson's equation. It only reverts to Poisson's equation when either $\partial\rho/\partial\phi = 0$ or $\phi = 0$, i.e., when ρ is constant or independent of ϕ, or when ϕ is constant. None of these cases is applicable to a nonuniformly doped semiconductor.

Conceptually, Equation (4.7) attempts to describe a system of fixed and mobile charges. It is, however, clearly impossible to maintain a nonzero charge distribution inside a conductor by means of an electrostatic field, and to maintain zero currents at the same time. If a net charge distribution is to be present at thermostatic equilibrium other than electrostatic forces must also be present. But the potential energies associated with those forces are missing in Equation (4.7). These potentials are the chemical potentials, μ_{cp} and μ_{cn}.

The rules of the variational formulation were violated when the solutions of Equation (3.17), $\mu_{cno} = e\phi_o$, and $\mu_{cpo} = -e\phi_o$, were

implicitly introduced into Equation (4.6). These solutions must result from variations of the functional and cannot be presupposed.

With these perceptions the stage is set for the derivation of the variational formulation for the carrier distributions. One simply must find all contributions to the density of the free energy.

Recalling Equations (3.19) and (3.20) and dropping the subscripts, o, for simplicity, one can write the charge distribution in terms of the chemical potentials as

$$\rho[\mu_{cp}(\vec{r}),\mu_{cn}(\vec{r}),\vec{r}] = e\{n_i\exp[\mu_{cp}/(kT)]-n_i\exp[\mu_{cn}/(kT)]+N(\vec{r}) \quad . \qquad (4.9)$$

The electrostatic energy in this system is still given by Equation (4.2), with Equation (4.9) describing $\rho(\vec{r})$. All of this energy contributes to the free energy. In addition there are contributions from the chemical energy, and these warrant careful consideration.

In Equation (4.9) the electron density is expressed as $n = n_i\exp[\mu_{cn}/(kT)]$. One recalls that this expression has been obtained by first approximating the Fermi function by the Boltzmann factor for describing the occupation probabilities of the quantum-mechanical electron states in the conduction band, and then summing up the resulting occupation numbers of all these states. This approximation is tantamount to treating the electrons as a quasi-classical gas; maintaining the indistinguishability, but disregarding the Pauli exclusion principle. This model must be applied consistently in deriving the chemical free energy.

Considering that the semiconductor region under investigation is in thermal contact with its environment, but cannot exchange

particles with it, the thermostatic equilibrium state is that with minimum Helmholtz free energy. One can then say with Feynman et al. [38] that the Helmholtz free energy in any subregion of the semiconductor must also be a minimum. The subregion, on the other hand, can exchange energy and particles with its neighbors, and it represents therefore a sample of a grand canonical ensemble. The density of the Helmholtz free energy thus results from the statistics of the grand canonical ensemble of independent, indistinguishable particles.

A heuristic argument yields the density of the chemical free energy, f_{cn}, in terms of the chemical potential, μ_{cn}, with little effort. In adding dn electrons to a unit volume with rigid but diathermic walls which already contains n electrons, the volume receives incremental work

$$\mu_n dn = dg_n + s_n dT - v dP = dg_n - v dP. \tag{4.10}$$

Performing this experiment over and over, starting with n = 0, one obtains

$$\int_0^n \mu_\nu d\nu = \int^g dg - v\int^P dP = g - vP = f_{cn}, \tag{4.11}$$

where g_n is the Gibbs free energy of the n particles in the unit volume, v. With $\mu_{cn} = (kT)\ln(n/n_i)$ the integral yields

$$f_{cn} = kT\int_0^n \ln(\nu/n_i)d\nu = (nkT)\ln(n/n_i) - nkT$$

$$= n_i \exp[\mu_{cn}/(kT)](\mu_{cn} - kT), \tag{4.12}$$

The term nkT is recognized as the term vP, and $(nkT)\ln(n/n_i) = n\mu_{cn} = g_n$. In deriving this result some liberties have been taken, and it seems appropriate to underpin the result more rigorously. To this end the Helmholtz free energy for a grand canonical ensemble of independent, indistinguishable particles will be calculated next. This energy is given by the equation

$$f_{cn} = (\bar{n}kT)\ln(z)-(kT)\ln[Q_g(z,v,T)], \tag{4.13}$$

where Q_g is the grand partition function; and where

$$z = \exp[\mu_{cn}/(kT)] = \bar{n}/n_i \tag{4.14}$$

It is noted that the particle density is now an expected value, as indicated by the bar above the symbol. The particle density can be obtained from the partition function by the differentiation

$$\bar{n} = z\partial\{\ln[Q_g(z,v,T)]\}/\partial z, \tag{4.15}$$

The grand partition function is related to the canonical partition function, Q_c, by

$$Q_g(z,v,T) = \sum_{n=0}^{\infty} z^n Q_{cn}(v,T). \tag{4.16}$$

Finally, for an ensemble of n independent, indistinguishable particles in the volume, v, at temperature, T, Q_{cn} is given by

$$Q_{cn} = (n!)^{-1}Q^n, \tag{4.17}$$

where Q is the single particle partition function,

$$Q = v(\sqrt{2\pi mkT}/h)^3 = v/\lambda_{Th}^3. \tag{4.18}$$

The single particle partition function can be interpreted as the ratio of the volume to the thermal volume of the particle, λ_{Th}^3, where λ_{Th} is the de Broglie wavelength of a particle with thermal velocity, v_{Th}. Inserting Equation (4.17) into (4.16) yields

$$Q_g(z,v,T) = \sum_{n=0}^{\infty} (n!)^{-1} z^n Q^n = \exp(zQ), \tag{4.19}$$

and inserting this expression into (4.15) leads to

$$\bar{n} = zQ = (v/\lambda_{Th})^3 \exp[\mu_{cn}/(kT)]. \tag{4.20}$$

In view of Equation (4.14) this further defines Q as

$$Q = n_i. \tag{4.21}$$

Inserting Equations (4.19) and (4.14) into (4.13), one obtains

$$f_{c\bar{n}} = \bar{n}\mu_{cn} - \bar{n}kT, \tag{4.22}$$

which is the same as Equation (4.12), if one only recognizes that the particle density in a subregion is an expected value also in (4.12), rather than a deterministic quantity.

Returning now to the variational problem, the total density of the Helmholtz free energy is, in view of Equations (4.2), (4.9), and (4.12),

$$w = e\phi\{n_i \exp[\mu_{cp}/(kT)] - n_i \exp[\mu_{cn}/(kT)] + N(\vec{r})\} - \tfrac{1}{2}\varepsilon(\nabla\phi)^2 +$$

$$+ n_i \exp[\mu_{cp}/(kT)](\mu_{cp} - kT) + n_i \exp[\mu_{cn}/(kT)](\mu_{cn} - kT), \tag{4.23}$$

the holes contributing in analogous fashion as the electrons. The functional contains now three different potential functions, ϕ, μ_{cp}, and μ_{cn}, and its variations yield three Euler-Lagrange equations. The variations with respect to μ_{cp} and μ_{cn} result in

$$n_i \exp[\mu_{cp}/(kT)][e\phi/(kT)+\mu_{cp}/(kT)+1-1] = 0, \tag{4.24}$$

$$n_i \exp[\mu_{cn}/(kT)][-e\phi/(kT)+\mu_{cn}/(kT)+1-1] = 0, \tag{4.25}$$

which specify the relations between the electrostatic and chemical potentials as

$$e\phi+\mu_{cp} = 0 = e\phi-\mu_{cn}, \tag{4.26}$$

The variation with respect to ϕ leads to

$$e\{n_i \exp[\mu_{cp}/(kT)]-n_i \exp[\mu_{cn}/(kT)]+N(\vec{r})\}+\nabla\cdot(\varepsilon\nabla\phi) = 0. \tag{4.27}$$

Finally, insertion of Equation (4.26) into (4.27) yields

$$e\{n_i \exp[-e\phi/(kT)]-n_i \exp[e\phi/(kT)]+N(\vec{r})\}+\nabla\cdot(\varepsilon\nabla\phi) = 0, \tag{4.28}$$

which is indeed Poisson's equation for doped semiconductor material. This demonstrates that the integral of Equation (4.23) is the correct functional.

Now, a few remarks about boundary conditions and related problems are in order. To this end the device structure of an insulated gate field effect transistor is considered as an example. The device layout and cross-section are sketched in Figure 4.1. A volume shaped essentially like a parallelepiped is delineated by the

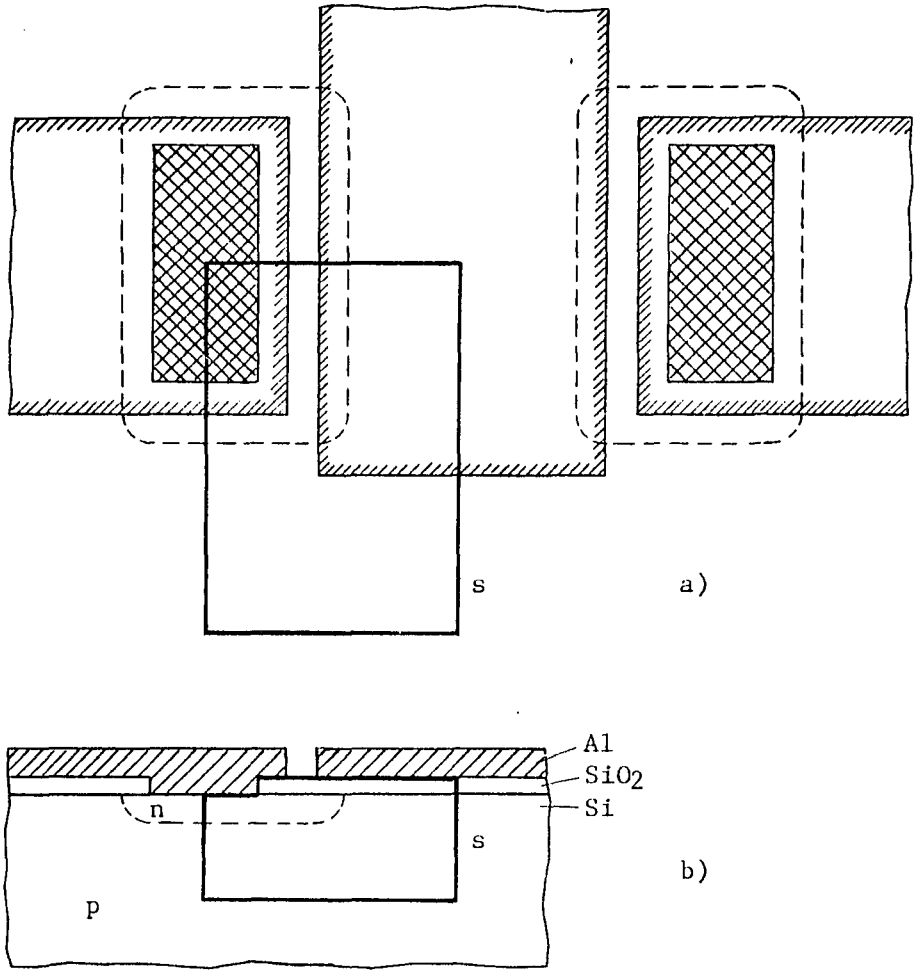

Figure 4.1. Geometry of an Insulated Gate Field Effect Transistor; (a) Lay-out; (b) Partial Cross-Section -- The aluminum electrodes are hatched, the contact holes are cross-hatched, and the metallurgical junctions are indicated by broken lines. The heavily drawn contours, s, indicate a parallelepiped-like region considered.

contour s. This volume contains a region of silicon and a region of silicon dioxide.

At the interface from the silicon to the silicon dioxide the dielectric permittivity changes from $12\varepsilon_o$ to $3.9\varepsilon_o$, and the band gap changes from 1.12 eV to 8 eV. The variational formulism has already been developed for a spatially varying permittivity. Recalling that the parameter n_i contains the information about the band structure in the silicon, one concludes that the different band structure in the silicon dioxide will merely be reflected in a different value of n_i for the oxide region. The formulation, however, is not restricted to a spatially constant n_i. The large band gap in the oxide makes the assumption that the carrier concentrations vanish in this region quite reasonable. Formally this can be achieved by letting n_i vanish in this region, which reduces the functional to the electrostatic energy alone, as a logical consequence of the absence of free carriers.

Fixed interface charges or volume charges in the insulator are of the same nature as the ionic charges in the silicon and are included in $N(\vec{r})$.

In conclusion, heterogeneous regions of semiconductors and insulators can be described by the variational formulation without complications.

The specification of boundary conditions for the thermostatic equilibrium state is simplified by Equation (4.26) which makes the specification of one of the three potentials sufficient. The potential ϕ is chosen because it is the only one also defined in the insulator,

The bottom face of the parallelepiped is chosen deep in the silicon, where the effects of the device structure are insignificant, and where only the features of the homogeneous substrate material remain. Assuming that the impurities are fully ionized, and that their concentration in the substrate is N_o, the potential is

$$\phi_o = \mu_{cn}/e = (kT/e)\ln(N_o/n_i). \tag{4.29}$$

This is the boundary condition for the bottom face of the region under investigation,

The four faces perpendicular to the surface of the silicon have been chosen parallel to the flow lines of the potentials. For the three faces which are partially underneath the metal electrodes this is true by virtue of symmetries in the structure. The fourth face is far enough away from the features of the device structure that its effects are again insignificant, and that the symmetry of the substrate causes the flow lines to be perpendicular to the surface of the silicon. Thus, condition (4.5) is satisfied on all four of the vertical faces. The remaining face was chosen along the interfaces with the aluminum electrodes and at the surface of the silicon dioxide. At thermostatic equilibrium the net electric charge contained in the device structure is zero. To avoid ambiguity, let all aluminum electrodes be in contact with one another so that all parts of the boundary at the metal interface are at the same potential. At the silicon-aluminum interface the energy difference between the conduction band edge of the silicon and the electrochemical potential (Fermi level) in the aluminum is equal to the barrier height, ϕ_B as was illustrated in Figure 3.4. From

that figure and from Equations (3.18) and (4.26) one concludes that the electrostatic and chemical potentials are

$$e\phi_{sm} = \mu_{cnsm} = -\mu_{cpsm} = \tfrac{1}{2}E_g - e\phi_B; \tag{4.30}$$

the index sm denoting semiconductor-metal interfaces. The barrier height is an empirical parameter which reflects the balance between the electron populations in the valence and conduction band states of the silicon on one side and in the conduction band states of the aluminum on the other.

The remainder of the boundary surface is the area of the top face outside the metal electrodes and is neither a surface of constant potential, nor is it spanned by flow lines. In fact, this surface has been introduced artificially in order to close the region under observation, while the device structure is truly open. The potential distribution of this part of the surface depends on the impurity profile in the silicon underneath as well as on external charges which may be present nearby. In order to fully specify the problem one must impose a reasonable potential distribution on this surface. As a guide one can consider that the electric field is normal to the silicon surface in regions remote from the metal electrodes and from the metallurgical junctions. In these regions the electric field must also be small, so that the potential approximately equals the surface potential of the silicon substrate,

$$\phi_s \simeq \phi_{si} = (kT/e)\ln(N_s/n_i); \tag{4.31}$$

si denoting the interface between the semiconductor and the insulator. Toward the metal edge the potential gradually changes from the value ϕ_s, given by Equation (4.31), to the value ϕ_{sm}, given by (4.30). The transition region is probably no wider than two times the thickness of the silicon dioxide layer.

This heuristic specification of the last part of the boundary condition may not be too scientifically pleasing. However, the carrier distributions are not very sensitive to this boundary condition, because the stray fields contain little energy, unless high charge concentrations exist nearby. Furthermore, if necessary one could determine the sensitivity of the carrier distributions to this part of the boundary conditions by calculating them for different potential distributions.

In concluding this section it should be pointed out that in the formal derivation of the variational principle some terms had to be introduced which are insignificant for all practical purposes. Yet, without the terms $kTn_i\exp[\mu_{cp}/(kT)]$, and $kTn_i\exp[\mu_{cn}/(kT)]$ in Equation (4.23) the formal solution could never be obtained in its correct form. As long as one does not have the exact solution, however, one does not know whether the missing parts are significant or not.

4.3 Variational Formulation of Carrier Transport

In Section 3.2 the carrier transport phenomena were reviewed from the point of view of thermodynamics. This was done primarily for defining variables and parameters, and for highlighting the assumptions behind the transport equations normally used in the semiconductor

device theory. Now, the thermodynamic description is resumed for further investigating the variational principle which governs carrier transport.

As mentioned earlier, the theory of irreversible thermodynamics postulates that the fluxes and generation rates of carriers and of heat adjust themselves such that the entropy generation rate is minimized. Transport theory for semiconductor devices is usually developed under the assumption of isothermal structures, and this simplification will also be made here. Further, the formulation will be restricted to stationary states of weak nonequilibrium.

The entropy generation rate per unit volume is expressed by Equation (3,11). For isothermal regions and stationary carrier flow this equation simplifies to

$$g_s = (1/T)[\vec{j}_p \cdot \nabla(-\mu_p) + \vec{j}_n \cdot \nabla(-\mu_n) - g_p \delta\mu_p - g_n \delta\mu_n], \tag{4.32}$$

because the net generation rates of neutral donors and ionized acceptors must vanish. Insertion of Equations (3,24), (3,25), and (3,29) into (4,32) leads to

$$g_s = (1/T)[(pm_p/e)(-\nabla\delta\mu_p)^2 + (nm_n/e)(-\nabla\delta\mu_n)^2 + \\ + (a/T)(\delta\mu_p + \delta\mu_n)^2], \tag{4.33}$$

if one considers that $\mu_{po} = \mu_{no} = 0$. Multiplying this equation by T translates from the entropy generation to heat generation or power-dissipation, $g_w = Tg_s$, and hence to the principle of minimum power dissipation. The total power dissipation, in terms of the deviations

of the electrical and chemical potentials from their equilibrium values, is given by the integral

$$G_w = \iiint_V \{(pm_p/e)[\nabla(\delta\mu_{cp}+e\delta\phi)]^2+(nm_n/e)[\nabla(\delta\mu_{cn}-e\delta\phi)]^2+$$

$$+(a/T)(\delta\mu_{cp}+\delta\mu_{cn})^2\}dr. \qquad (4.34)$$

Consider sufficiently small deviations from equilibrium, that the concentrations p and n can be approximated by their equilibrium values, p_o and n_o. Then the integrand of Equation (4.34) becomes strictly a quadratic function of the potentials and their gradients. The variations with respect to $\delta\phi$, $\delta\mu_{cp}$, and $\delta\mu_{cn}$ then lead to

$$0 = \nabla\cdot[2e(p_o m_p/e)\nabla(\delta\mu_{cp}+e\delta\phi)-2e(n_o m_n/e)\nabla(\delta\mu_{cn}-e\delta\phi)$$

$$= 2e\nabla\cdot(\vec{j}_p-\vec{j}_n) = 2\nabla\cdot\vec{j}_{el}, \qquad (4.35)$$

$$0 = 2(a/T)(\delta\mu_{cp}+\delta\mu_{cn})-\nabla\cdot[2(p_o m_p/e)\nabla(\delta\mu_{cp}+e\delta\phi)$$

$$= 2(g_p-\nabla\cdot\vec{j}_p), \qquad (4.36)$$

$$0 = 2(a/T)(\delta\mu_{cp}+\delta\mu_{cn})-\nabla\cdot[2(n_o m_n/e)\nabla(\delta\mu_{cn}-e\delta\phi)$$

$$= 2(g_n-\nabla\cdot\vec{j}_n). \qquad (4.37)$$

These are the continuity equations of the stationary electric current density, j_{el}, and of the flux densities of holes and electrons, j_p and j_n.

Unfortunately the approximation of p and n by p_o and n_o is only valid for uselessly small deviations from the equilibrium state. Since

the carrier densities are given by

$$p = p_o \exp[\delta\mu_{cp}/(kT)], \quad n = n_o \exp[\delta\mu_{cn}/(kT)], \tag{4.38}$$

the approximations are only good if $\delta\mu_{cp}$, $\delta\mu_{cn} \ll kT$. In a device at room temperature these conditions are only met when all applied voltages are substantially less than 25 mV. They highlight the limitations inherent in the theory of irreversible thermodynamics which is a linear theory.

The formulation only contains the constant term in a Taylor expansion of any of the transport coefficients about the equilibrium state in terms of the electrical and chemical potentials. The functional dependences of the transport coefficients on these potentials are not contained in the theory. One may conjecture that only a self-consistent derivation of the pair generation rate, the carrier mobility and the carrier diffusion coefficient, as functions of the potentials, from a collision model may lead to the formulation of an analytically provable, more general formulation of a variational principle for carrier transport. Short of such a proven formulation the application of the finite element method to problems of carrier transport in solid state devices rests on weak ground.

There can be no doubt that the current and potential distributions under stationary operating conditions are indeed governed by a variational principle. Ordinary device structures exhibit wide operating regions in which the current response is uniquely determined by the applied steady state potentials. Restoring forces in a general sense must, therefore, exist, which establish and maintain the

appropriate internal potential and flux distributions and make any spontaneous deviation from them subside. These forces must just vanish when the proper distributions are reached. The forces can be considered to be gradients of generalized potentials, and these potentials must be stationary when the equilibrium flux patterns are established. Again, these generalized potential functions must be extremal in each subregion of the device, and hence their integrals over the total region must also be extremal; the sum of the integrals being the desired functional.

The derivation and analytical proof of a variational principle for strong nonequilibria will not be pursued further in this dissertation.

CHAPTER 5

REVIEW, CONCLUSIONS, OUTLOOK

The major results of this work fall into two areas of modeling of integrated devices; calculation of one-, two-, and three-dimensional impurity profiles, and calculation of carrier distributions in the device structures. The efforts have been directed toward formulating more accurate and detailed, but still simple, descriptions of the distributions of impurities and carriers. A guiding principle has been the achievement of a consistent level of detail and accuracy, commensurate with the capabilities of modern device processing technology, in the description of the individual properties of a device structure.

The dissertation not only outlines the successful techniques, but also discusses and explains the inadequacies of unsuccessful methods. Since the calculation of impurity profiles or carrier distributions, whether numerically or by analytic approximations, requires extensive and time-consuming preparations, it is very useful to be aware of those methods which will not yield acceptable results.

Considerable space has been devoted to one-dimensional cases, especially in the second chapter. This has mainly two reasons. First, many aspects of the theory can be demonstrated more clearly by means of one-dimensional formulations without loss of generality. The second reason, particularly valid for numerical evaluations, is

economy in data processing. For example, the calculations for the three-dimensional profiles illustrated in Figures 2.4 and 2.5 required 40 s per node on an IBM 360-44 computer.

For the figures the variables were calculated at about 600 nodes, while the computation of the entire profile would involve about 6000 nodes.

In summary, the research activities leading to this dissertation have resulted in a number of contributions to the device modeling technology. The techniques for modeling impurity profiles of integrated device structures have been advanced by four distinct innovations.

1. Simple analytic approximations have been formulated which describe the impurity profiles of two-step diffusions without reoxidation more accurately and over a wider range of the concentration than all previous analytic and numerical methods. The theory of the new approximations allows the adaptation of the accuracy to the requirements of the situation by adjusting the number of terms in the formulae.
2. A semi-analytic technique, based on Green's functions has been derived for computing three-dimensional impurity profiles of predeposition diffusions through rectangular windows in the diffusion mask. This is the only known technique today with which such profiles can be calculated over a sufficient range of the concentration with uniform and known accuracy. The ability to accurately calculate such profiles is a prerequisite to the development of approximations, because they represent

the only standards against which the quality of the approximations can be measured.

3. A simple analytic approximation technique has been formulated which describes three-dimensional predeposition profiles from rectangular windows. The complexity of these approximations can also easily be adapted to the accuracy requirements. The formulation can be simplified to the description of two-dimensional profiles from strip-like diffusion windows in a natural and simple way.
4. The known formulism which allows the dependence of the diffusivity on impurity concentration to be included in the calculation of one-dimensional profiles has been extended to the three-dimensional formulation.

There are five contributions to the techniques for modeling the behavior of the carriers in integrated devices.

1. The existing theories of effects of high impurity concentrations have successfully been merged with the conventional, simple semiconductor device theory, thereby extending the applicability of this theory to the high concentrations frequently used in modern integrated devices. This improvement has been achieved in analytic formulation with very minor complications of the simple formulation.
2. The close relationship suspected to exist between the carrier concentrations in uniformly doped material and in highly nonuniform regions in semiconductor device structures has been

proven analytically. This relationship holds for distinctly extrinsic regions, except next to the metallurgical junctions, contacts and surfaces.

3. A variational principle which governs the carrier distribution in an arbitrary semiconductor region at thermostatic equilibrium has been formulated from the basic principles and postulates of physics. Its correctness has been proven analytically. This variational principle can be used for calculating equilibrium carrier distributions in those semiconductor regions, which are not amenable to analytic approximations, by means of the finite element method.
4. A variational formulation of carrier transport has been derived from the principles of irreversible thermodynamics. Its validity has also been proved by analytically deriving the transport equations from the functional. However, the formulation illustrates that the technique is limited to very weak nonequilibrium and can only model operating conditions with applied electrical potential differences of less than 25 mV.
5. The derivations concerning carrier distributions and transport have been formulated in the nomenclature of thermodynamics. This terminology has been defined in Section 3.2. It provides more clarity in the formulations of nonequilibrium processes in semiconductor devices.

The achievements of this work have made numerous parts of the semiconductor device models compatible with modern devices in accuracy

and detail. However, not all problems have been solved. Two crucial problems remain unresolved.

1. No adequate formulation, analytic or otherwise, exists to date which would accurately describe the impurity profiles resulting from two-step diffusions with reoxidation during the drive-in step.
2. No practically useful method is yet available for the numerical solution of three-dimensional transport problems. Even two- and one-dimensional problems are very cumbersome to solve with present methods.

The first problem cannot be adequately solved with the approximation techniques developed in Chapter 2. The accuracy of published solutions is no better and is also totally insufficient. The problem, moreover, is of a nature which does not leave much hope for success with numerical methods.

The numerical solutions of transport problems have, so far, been carried out by simulating the governing differential equations with finite difference algorithms. Solved problems are usually one-dimensional, rarely two-dimensional and never three-dimensional. The solutions also contain simplifications and linearizations which are not tolerable in modeling of modern integrated devices. The simulation of carrier transport in these devices is out of reach of the finite difference techniques. The finite element method appears much more promising, but so far no variational formulation for general carrier transport has been established and proved analytically in the rigorous

manner postulated earlier. The development of this variational principle will entail a careful reconsideration of many aspects, approximations and simplifications inherent in the solid state theory normally used in integrated electronics. Even some of the physical concepts behind this theory may have to be extended. Such investigations will, first of all, distinctly contribute to a much better understanding of the concepts of carrier transport in nonuniformly doped regions and of carrier behavior in general. It may also lead to advances in the theory of nonequilibrium thermodynamics. Last, but not least, the variational formulation will allow the simulation of carrier transport by the finite element method, which is expected to open the way to new approximation methods in this field.

The successful solution of the above two problems, in combination with the solved problems listed and with the systematic modeling theory [6, 7], will lead to an overall modeling technique capable of describing modern integrated devices with sufficient accuracy. Such an achievement will close the last gaps in an automatic integrated circuit design system as envisioned at the very beginning of this dissertation.

REFERENCES

1. Cornell University, School of Electrical Engineering, System Theory Research Group, "FORTRAN Computer Routine CORNAP," 1968.

2. Weeks, W. T., A. J. Jimenez, G. W. Mahoney, D. Mehta, H. Qassemzadeh, and T. R. Scott, "Algorithms for ASTAP--A Network Analysis Program," IEEE Trans. on Circuit Theory, Vol. CT-20, pp. 628-634, 1973.

3. Linvill, J. G., and J. F. Gibbons, Transistors and Active Circuits, McGraw-Hill, New York, 1961.

4. Lindholm, F. A., "Device Modeling for Computer-aided Analysis and Design of Integrated Circuits," Solid State Electronics, Vol. 12, pp. 831-840, 1969.

5. Lindholm, F. A., and D. J. Hamilton, "A Systematic Modeling Theory for Solid State Devices," Solid State Electronics, Vol. 7, pp. 771-783, 1964.

6. Hamilton, D. J., F. A. Lindholm, and A. H. Marshak, Principles and Applications of Semiconductor Device Modeling, Holt, Rinehart, and Winston, New York, 1971.

7. Fossum, J. G., Systematic Computer-aided Three-dimensional Modeling of Integrated Bipolar Devices, Ph.D. Dissertation, University of Arizona, Tucson, Arizona, 1971.

8. Grove, A. S., Physics and Technology of Semiconductor Devices, John Wiley and Sons, New York, 1967.

9. Morse, P. M., and H. Feshbach, Methods of Theoretical Physics, Part I, McGraw-Hill, New York, 1953.

10. Oden, J. T., Finite Elements of Nonlinear Continua, McGraw-Hill, New York, 1972.

11. Zienkiewicz, O. C., The Finite Element Method in Engineering Science, McGraw-Hill, London, 1971.

12. Gajda, W. J., and J. H. Jackson, "The Gaussian Planar PN Junction," Proc. Nat'l Electronics Conf., Vol. 26, pp. 234-237, 1970.

13. Carslaw, H. S., and J. C. Jaeger, Conduction of Heat in Solids, Clarendon Press, Oxford, 1959.

14. Kennedy, D. P., Mathematical Investigations of Semiconductor Devices, Report AFCRL-66-358, Air Force Cambridge Laboratories, Bedford, Mass., 1966.

15. Abramowitz, M., and I. A. Stegun (eds.), Handbook of Mathematical Functions, U.S. Dept. of Commerce, Nat. Bureau of Standards, Washington, D.C., 1968.

16. Huang, J. S. T., and L. C. Welliver, "On the Redistribution of Boron in the Diffused Layer During Thermal Oxidation," J. Electrochem. Soc., Vol. 117, pp. 1577-1580, 1970.

17. Carnahan, B., H. A. Luther, and J. D. Wilkes, Applied Numerical Methods, John Wiley and Sons, New York, 1969.

18. Gurtin, M. E., "Variational Principles for Linear Initial Value Problems," Quarterly of Applied Mathematics, Vol. 12, pp. 252-256, 1964.

19. Stackgold, I., Boundary Value Problems of Mathematical Physics, Macmillan, New York, 1968.

20. Fourier, J., The Analytical Theory of Heat, Dover, New York, 1955; translated from: Fourier, J. B. J., Théorie Analytique de la Chaleur, 1822.

21. Kurtz, A. D., and R. Yee, "Diffusion of Boron into Silicon," J. Appl. Physics, Vol. 31, pp. 303-305, 1960.

22. Lehovec, K., and A. Slobodskoy, "Diffusion of Charged Particles into a Semiconductor under Consideration of the Built-in Field," Solid State Electronics, Vol. 3, pp. 45-50, 1961.

23. Smith, A. C., J. F. Janak, and R. B. Adler, Electronic Conduction in Solids, McGraw-Hill, New York, 1967.

24. Keyes, R. W., "Effect of Randomness in the Distribution of Impurity Ions on FET Thresholds in Integrated Electronics," IEEE J. Solid-State Circuits, Vol. SC-10, pp. 245-247, 1975.

25. Sze, S. M., Physics of Semiconductor Devices, John Wiley and Sons, New York, 1969.

26. DeMan, H. J., "The Influence of Heavy Doping on the Emitter Efficiency of a Bipolar Transistor," IEEE Trans. Electron Dev., Vol. ED-18, pp. 833-835, 1971.

27. Parmenter, R. H., "Energy Levels of a Disordered Alloy," Phys. Rev., Vol. 97, pp. 587-598, 1955.

28. Parmenter, R. H., "Energy Levels of a Disordered Alloy," Phys. Rev., Vol. 104, pp. 22-32, 1956.

29. Kane, E. O., "Thomas-Fermi Approach to Impure Semiconductor Band Structure," Phys. Rev., Vol. 131, pp. 79-88, 1963.

30. Morgan, T. N., "Broadening of Impurity Bands in Heavily Doped Semiconductors," Phys. Rev., Vol. 139, pp. A343-A348, 1965.

31. Blakemore, J. S., Semiconductor Statistics, Pergamon Press, 1962.

32. Van Overstraeten, R. J., H. J. DeMan, and R. P. Mertens, "Transport Equations in Heavy Doped Silicon," IEEE Trans. Electron Dev., Vol. ED-20, pp. 290-298, 1973.

33. Kleppinger, D. D., An Extension of the Engineering Theory of Semiconductors with Applications, Ph.D. Dissertation, University of Florida, Gainsville, Florida, 1970.

34. Buturla, E. M., and P. E. Cottrell, "Two-dimensional Finite Element Analysis of Semiconductor Steady State Transport Equations," Internat. Conf. Comput. Methods Nonlin. Mech., Austin, Texas, 1974.

35. Cottrell, P. E., and E. M. Buturla, "Steady State Analysis of Field Effect Transistors via the Finite Element Method," Internat. Electron Dev. Conf., Washington, D.C., 1975.

36. Maxwell, J. C., A Treatise on Electricity and Magnetism, Vol. 1, Dover, New York, 1954.

37. Millar, W., "Some General Theorems for Non-Linear Systems Possessing Resistance," Phil. Mag., Ser. 7, pp. 1150-1160, 1951.

38. Feynman, R. P., R. B. Leighton, and M. Sands, The Feynman Lectures on Physics, Addison-Wesley, Reading, Mass., 1965.