Implications of Analytical Investigations About the Semiconductor Equations on Device Modeling Programs

SIEGFRIED SELBERHERR, MEMBER, IEEE, AND CHRISTIAN A. RINGHOFER

Abstract—This paper gives guidelines for the development of computer programs for the numerical simulation of semiconductor devices. For this purpose, the basic mathematical results on the corresponding elliptic boundary value problem are reviewed. Particularly, existence, smoothness, and structure of the solutions of the fundamental semiconductor equations are discussed. Various feasible approaches to the numerical solution of the semiconductor equations are described. Much emphasis is placed on constructive remarks to help authors of device simulation programs make decisions on their code design problems. Thus criteria for an optimal mesh generation strategy are given. The iterative solution of the systems of nonlinear and linear equations obtained by discretizing the semiconductor equations is discussed. An example shows the power of these concepts combined with modern numerical methods in comparison to classical approaches.

I. Introduction

THE CHARACTERISTIC feature of early device modeling is the separation of the interior of the device into different regions, the treatment of which could be simplified by various assumptions like special doping profiles, complete depletion, and quasineutrality. These separately treated regions were simply put together to produce the overall solution. If results in an analytically closed form are intended, any other approach is prohibitive. Fully numerical modeling based on partial differential equations [61] which describe all different regions of semiconductor devices in one unified manner was first suggested by Gummel [29] for the one-dimensional bipolar transistor. This approach was further developed and applied to p-n-junction theory by De Mari [13], [14] and to IMPATT diodes by Scharfetter and Gummel [50].

A two-dimensional numerical analysis of a semiconductor device was first carried out by Kennedy and O'Brien [35] who investigated the junction field effect transistor. Since then, two-dimensional modeling has been applied to all important semiconductor devices. There are so many papers of excellent repute that it would be unfair to cite only a few. Additionally, the first results on three-dimensional device modeling have now been published. Time dependence has been investigated by [37] and [44] and models in three space dimensions have been announced by, e.g., [8], [11], [67], and [68].

In spite of all these important and successful activities, the need for economic and highly user-oriented computer programs becomes more and more apparent in the field of device modeling. Especially for MOS devices, which have evolved since their invention by Kahng and Atalla [32] to an incredible standard, modeling in two space dimensions has become inherently important because current flow controlled by a perpendicular field is an intrinsically two-dimensional problem. One such program which has been applied successfully in many laboratories is called CADDET [59]. We have also tried to bridge that gap and developed MINIMOS [51], [53] for the two-dimensional static analysis of planar MOS transistors.

II. Analysis of the Static Semiconductor Equations

In this section, we reivew some of the existing analytical results for the fundamental semiconductor equations concerning existence and structure of their solutions. These results are of importance in both the theoretical and practical context, since—as we will see in the next section—the knowledge of the structure and smoothness properties of solutions is indeed essential for the development of a numerical solution method. The most familiar model of carrier transport in a semiconductor device has been proposed by Van Roosbroeck [61]. It consists of Poisson's equation (2.1), the current continuity equations for electrons (2.2) and holes (2.3), and the current relations for electrons (2.4) and holes (2.5)

$$\operatorname{div} \epsilon \cdot \operatorname{grad} \psi = -q \cdot (p - n + C) \tag{2.1}$$

$$\operatorname{div} \vec{J}_n = -q \cdot R \tag{2.2}$$

$$\operatorname{div} \vec{J}_p = q \cdot R \tag{2.3}$$

$$\vec{J}_n = -q \cdot (\mu_n \cdot n \cdot \operatorname{grad} \psi - D_n \cdot \operatorname{grad} n) \tag{2.4}$$

$$\vec{J}_p = -q \cdot (\mu_p \cdot p \cdot \operatorname{grad} \psi + D_p \cdot \operatorname{grad} p). \tag{2.5}$$

The relations form a system of coupled partial differential equations. Poisson's equation, coming from Maxwell's laws, describes the charge distribution in the interior of a semiconductor device. The balance of sinks and sources for electron and hole currents is characterized by the continuity equations. The current relations describe the absolute value, di-

Manuscript received August 8, 1983. This work is supported by the "Fonds zur Förderung der wissenschaftlichen Forschung" Project S22/11, by the Research Laboratory of Siemens AG, Munich, Germany, and the U.S. Army under Contract DAAG-29-80-C-0041.

S. Selberherr is with Abteilung Physikalische Elektronik, Institut fuer Allgemeine Elektrotechnik und Elektronik, TU Wien, A-1040 Wien, Austria.

C. Ringhofer is with the Mathematics Research Center, University of Wisconsin, Madison, WI 53706.

0028-0070/84/0100-0052$01.00 © 1984 IEEE

rection, and orientation of electron and hole currents. The continuity equations and the current relations can be derived from Boltzmann's equation by not at all trivial means. It is not our intention to present in this paper the ideas behind these considerations. The interested reader is refered to [61] and its secondary literature or textbooks on semiconductor physics e.g., [7], [31], [52], [56].

2.1. The Validity of the Basic Semiconductor Equations

It is of prime importance to be aware that (2.4) and (2.5) are not capable of exactly describing all phenomena occurring in real devices. For instance, they do not characterize effects which are caused by degenerate semiconductors (e.g., heavy doping). References [38], [60], and [63] discuss some modifications of the current relations, which partially take into account the consemences introduced by degenerate semiconductors (e.g, invalidity of Boltzmann's statistics, bandgap narrowing). These modifications are not at all simple and lead to problems especially in the formulation of boundary conditions [47], [62]. In case of modeling MOS devices, degeneracy, owing to the relatively low doping in the channel region is practically irrelevant. For modern bipolar devices, though, bearing in mind shallow and extraordinarily heavily doped emitters, it is an absolute necessary to account for local degeneracy of the semiconductor.

Just as further examples, (2.4) and (2.5) do not describe velocity overshoot phenomena which become apparent at feature lengths of 0.1 μm for silicon and 1 μm for gallium-arsenide [25]. Certainly no effects which are due to ballistic transport (the existence of which is still questionable [30]) are included. The latter start to become important for feature sizes below 0.01 μm for silicon and 0.1 μm for gallium-arsenide [26]. Considering the state of the art of device miniaturization, neither effect has to bother the modelists of silicon devices. For gallium-arsenide devices new ideas are mandatory in the near future [25], [45], [46].

2.2. Domain and Boundary Conditions

Most of the existing programs which solve the semiconductor equations are restricted to a rectangular device geometry. This is not essential as far as the analysis of the equations is concerned. In this chapter we shall assume that (2.1)-(2.5) are posed in a domain D of $\mathbb{R}^n$ $(n = 1, 2, 3)$ with a piecewise smooth boundary ∂D. Equations (2.1)-(2.5) are subject to a mixed set of Dirichlet and Neumann boundary conditions. This means that ∂D consists of three parts, $\partial D = \partial D_1 \cup \partial D_2 \cup \partial D_3$. ∂D_1 denotes the part of the boundary where the device is surrounded by insulating material. There one assumes the boundary conditions

$$\partial\Psi/\partial\vec{n}_\perp = \partial n/\partial\vec{n}_\perp = \partial p/\partial\vec{n}_\perp = 0. \tag{2.6}$$

Here $\vec{n}_\perp$ denotes the unit normal vector on ∂D which exists everywhere except at a finite number of points (arbitrarily defined corners of the simulation geometry). ∂D_2 denotes the part of the boundary corresponding to the ohmic contacts. There Ψ, n, and p are prescribed. The boundary conditions can be derived from the applied bias Ψ_D and the assumptions of thermal equilibrium and vanishing space charge

$$\Psi = \Psi_D + \Psi_{\text{built in}} \qquad n \cdot p = n_i^2 \quad \text{and} \quad n - p - C = 0. \tag{2.7}$$

The last two conditions in (2.7) can be rewritten as

$$n = (\sqrt{C^2 + 4 \cdot n_i^2} + C)/2$$
$$p = (\sqrt{C^2 + 4 \cdot n_i^2 - C})/2. \tag{2.8}$$

In many applications it is desired to consider controlled insulator-semiconductor interfaces (e.g., MOS devices). So, ∂D_3 denotes the part of the boundary which corresponds to such an interface. There we have the interface conditions

$$\vec{J}_n \cdot \vec{n}_\perp = \vec{J}_p \cdot \vec{n}_\perp = 0$$
$$\epsilon_{\text{sem}} \cdot \partial\Psi/\partial\vec{n}_\perp\big|_{\text{sem}} = \epsilon_{\text{ins}} \cdot \partial\Psi/\partial\vec{n}_\perp\big|_{\text{ins}}. \tag{2.9}$$

Again $\vec{n}_\perp$ denotes the normal vector on ∂D. ϵ_{sem} and ϵ_{ins} denote the permittivity constants for the semiconductor and the insulator, respectively. $\partial\Psi/\partial\vec{n}_\perp|_{\text{sem}}$ and $\partial\Psi/\partial\vec{n}_\perp|_{\text{ins}}$ denote the one-sided limits of the derivatives perpendicular to the interface approaching the interface. Within the insulator, the Laplace equation: div grad $\Psi = 0$ holds.

2.3. Dependent Variables

For analytical purposes it is often useful to use other variables than n and p to describe the system (2.1)-(2.5). Two other sets of variables which are frequently employed are $(\Psi, \varphi_n, \varphi_p)$ and (Ψ, u, v) which relate to the set (Ψ, n, p) by

$$n = n_i \cdot e^{(\Psi - \varphi_n)/U_t}, \qquad p = n_i \cdot e^{(\varphi_p - \Psi)/U_t} \tag{2.10}$$

$$n = n_i \cdot e^{\Psi/U_t} \cdot u, \qquad p = n_i \cdot e^{-\Psi/U_t} \cdot v. \tag{2.11}$$

Equation (2.10) can be physically interpreted as the application of Boltzmann statistics. However, (2.10) also can be regarded as a purely mathematical change of variables so that the question of the validity of the Boltzmann statistics does not need to be considered. The use of $(\Psi, \varphi_n, \varphi_p)$ *a priori* excludes negative carrier densities n and p, which may be present as undesired nonphysical solutions of (2.1)-(2.5) if we use (Ψ, n, p) or (Ψ, u, v) as dependent variables. As we will see later in this section, the advantage of the set (Ψ, u, v) is that the continuity equations (2.2), (2.3) and current relations (2.4), (2.5) become self adjoint. This also has an important impact on the use of iterative schemes for the solution of the evolving linear systems (cf. Section IV). However, owing to the enormous range of the values of u and v, the sets (Ψ, n, p) or $(\Psi, \varphi_n, \varphi_p)$ have to be preferred for actual computations. We personally favor the set (Ψ, n, p).

2.4. The Existence of Solutions and Scaling

The basic answer to the question of existence of solutions can be found in Mock [43] or under slightly different assumptions in Bank *et al.*, [5]. Both proofs are based on Schauder's fixpoint theorem. They are both valid for arbitrarily shaped domains and boundary conditions of the type previously described without an interface ($\partial D_3 = \{ \}$). Both papers consider the case of vanishing generation/recombination rate ($R = 0$ in (2.2), (2.3)). In the setting of Mock, (Ψ, u, v) are used as dependent variables. The equations are scaled so that the intrinsic carrier density n_i, the thermal voltage U_t, and the ratio elementary charge/permittivity are equal to unity. Thus combining the continuity equations (2.2), (2.3) and current relations (2.4), (2.5), we have the system

$$\text{div grad } \Psi = e^{\Psi} \cdot u - e^{-\Psi} \cdot v - C \tag{2.12}$$

$$\text{div } (e^{\Psi} \cdot \text{grad } u) = 0 \tag{2.13}$$

$$\text{div } (e^{-\Psi} \cdot \text{grad } v) = 0. \tag{2.14}$$

Then a map $M: \Psi \rightarrow y$ is defined (details in [4] or [43]) such that the evaluation of M requires the solution of (2.13) and (2.14). A fixpoint ψ^* of M (M(W*) = W*) together with the according functions (u, v) is a solution of the whole system (2.12)-(2.14). The exitence of a fixpoint is shown by Schauder's fixpoint theorem. Questions concerning the degree of smoothness of these solutions (the existence of derivatives) are discussed in [42].

However, Schauder's theorem is not constructive and does not indicate that iterating the map M will actually lead to the fixpoint. Moreover, it does not give any information about the structure of the solution which is of vital interest for actual computations. Since the dependent variables in the system (2.1)-(2.5) are of different order of magnitude and show a strongly different behavior in regions with small and large space charge, the first step towards a structural analysis of (2.1)-(2.5) has to be an appropriate scaling. A standard way of scaling (2.1)-(2.5) has been given by De Mari [14]. There Ψ is scaled by the thermal voltage U_t, n and p are scaled by n_i (similar to Mock [43]), and the independent variables are scaled such that all multiplying constants in Poisson's equation become unity. Although physically reasonable, this approach has the disadvantage that n and p in general are still several orders of magnitude larger than Ψ. A scaling which reduces Ψ, n, and p to the same order of magnitude has been given by Vasiliev'a and Butuzov [65]. This approach makes the system (2.1)-(2.5) accessible to an asymptotic analysis which is given together with applications in [39]-[41]. There, n and p are scaled by the maximum absolute value of the net doping C and the independent variables are scaled by the characteristic length of the device. More precisely, the following scaling factors are employed:

quantity	symbol	value
$\vec{x}$	l	$\max(\vec{x} - \vec{y}),\ \vec{x}, \vec{y}$ in D
Ψ	U_t	$k \cdot T/q$
n, p	α	$\max \lvert C \rvert$

(2.15)

After scaling, the equations become

$$\lambda^2 \cdot \text{div grad } \Psi = n - p - C$$

$$\text{div (grad } n - n \cdot \text{grad } \Psi) = -R$$

$$\text{div (grad } p + p \cdot \text{grad } \Psi) = -R. \tag{2.16}$$

Here, for simplicity only, μ_n and μ_p have been assumed to be constant. It should be noted that the following analysis also holds if the usual smooth dependence of μ_n and μ_p on n, p, and grad Ψ, e.g., [54], is assumed. Since the independent variable $\vec{x}$ has been scaled, (2.16) is now posed on a domain D^s with maximal diameter equal to 1. The small constant λ^2 multiplying the Laplacian in (2.16) is the minimal Debye length of the device

$$\lambda^2 = \frac{\epsilon \cdot U_t}{l^2 \cdot q \cdot \alpha} \tag{2.17}$$

l and α are defined in (2.15). Thus for high doping ($\alpha >> 1$), λ^2 will be small. For instance, for a silicon device with characteristic length 25 μm and $\alpha = 10^{20}$ cm^{-3}, we compute for λ^2 at approximate room temperature $T = 300$ K: $\lambda^2 = 4 \times 10^{-10}$.

R denotes again the scaled generation/recombination rate. In the analysis given in [41], the usual Shockley-Read-Hall term has been used, which after scaling is of the form

$$R = \frac{n \cdot p - (\zeta\lambda)^4}{n + p + 2 \cdot (\zeta\lambda)^2}, \qquad \zeta = \frac{1}{2} \tag{2.18}$$

R is in general a (not necessarily mildly) nonlinear function of n, p, and grad Ψ. Thus different models of R may influence the analytical results requite drastically. This is obviously to be expected, because in many operating conditions the device behavior depends strongly on the net generation/recombination R.

2.5. The Singular Perturbation Approach

Equation (2.16) represents a singularly perturbed elliptic system with perturbation parameter λ. The advantage of this interpretation is that we can now obtain information about the structure of solutions (2.16) by using asymptotic expansions. In the subdomains of D^s where the solutions behave smoothly, we expand them into power series of the form

$$w(\vec{x}, \lambda) = \sum_{i=0}^{\infty} w_i^{\sim}(\vec{x}) \cdot \lambda^i, \qquad w = (\Psi, n, p)^T \tag{2.19}$$

which implies a smooth dependence on λ. C—the scaled doping—is smooth in these subdomains and exhibits a sharp transition across the p-n junctions in the device. For the case of an abrupt junction this behavior is represented by a discontinuity across an $n - 1$ dimensional manifold Γ: ($x = x(s)$, s of R^{n-1}) in the device. Thus Γ is a point in one dimension, a curve in two dimensions and a surface in three dimensions. Of course, one curve or surface has to be used for each junction. Since the procedure is the same for each of the junctions, it is demonstrated only for one junction. In the case of an exponentially graded doping profile, C consists of two parts

$$C = C^{\sim} + C^{\wedge} \tag{2.20}$$

where $C^{\sim}$ and $C^{\wedge}$ are discontinous, $C^{\sim}$ is piecewise smooth, and $C^{\wedge}$ is exponentially decaying to zero away from Γ. In the vicinity of Γ, the expansion (2.19) is not valid and has to be supplemented by a "layer" term according to the singular perturbation analysis

$$w(\vec{x}, \lambda) = \sum_{i=0}^{\infty} [w_i^{\sim}(\vec{x}) + w_i^{\wedge}(s, t/\lambda)] \cdot \lambda^i,$$

$$w = (\Psi, n, p)^T. \tag{2.21}$$

Here the following coordinate transformation has been employed. For a point in the vicinity of Γ, s denotes the parameter value at the nearest point on Γ and t denotes its distance perpendicular to Γ (cf. Fig. 1). Thus the solution of the semiconductor equations exhibits internal layers at p-n junctions.

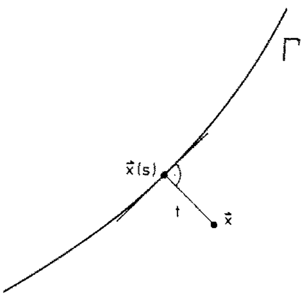


Fig. 1. Local coordinates of the layer solution.

The $w_i^{\sim}$ and $w_i^{\wedge}$ in (2.21) can now be determined separately and the structure of the solution is given by its partition into the smooth part $\Sigma\, w_i^{\sim} \cdot \lambda^i$ and its rapidly varying part $\Sigma\, w_i^{\wedge} \cdot \lambda^i$. $w_0^{\sim}$ has to satisfy the reduced equations

$$0 = n_0^{\sim} - p_0^{\sim} - C^{\sim} \tag{2.22}$$

$$\text{div}\,(\text{grad}\, n_0^{\sim} - n_0^{\sim} \cdot \text{grad}\, \Psi_0^{\sim}) = -R^{\sim} \tag{2.23}$$

$$\text{div}\,(\text{grad}\, p_0^{\sim} + p_0^{\sim} \cdot \text{grad}\, \Psi_0^{\sim}) = -R^{\sim}. \tag{2.24}$$

For the sake of simplicity, but without loss of generality, the mobilities μ_n and μ_p have been assumed to be constant. Equations (2.22)-(2.24) subject to the boundary conditions (2.6)-(2.9). Of course, the condition of vanishing space charge is redundant with (2.22). Since $C^{\sim}$ is discontinous at Γ and (2.22)-(2.24) represent a second-order system of two equations, four "interface conditions" have to be imposed at Γ. They are of the form

$$n_0^{\sim} \cdot e^{-\Psi_0^{\sim}}\Big|_{\vec{x}=\vec{x}-} = n_0^{\sim} \cdot e^{-\Psi_0^{\sim}}\Big|_{\vec{x}=\vec{x}+} \tag{2.25}$$

$$p_0^{\sim} \cdot e^{\Psi_0^{\sim}}\Big|_{\vec{x}=\vec{x}-} = p_0^{\sim} \cdot e^{\Psi_0^{\sim}}\Big|_{\vec{x}=\vec{x}+} \tag{2.26}$$

$$\vec{J}_{n_0}^{\sim} \cdot n_\perp\Big|_{\vec{x}=\vec{x}-} = \vec{J}_{n_0}^{\sim} \cdot n_\perp\Big|_{\vec{x}=\vec{x}+} \tag{2.27}$$

$$\vec{J}_{p_0}^{\sim} \cdot n_\perp\Big|_{\vec{x}=\vec{x}-} = \vec{J}_{p_0}^{\sim} \cdot n_\perp\Big|_{\vec{x}=\vec{x}+} \tag{2.28}$$

where $w|_{\vec{x}-}$ and $w|_{\vec{x}+}$ denote the one-sided limits of w as x tends to Γ from each side. $\vec{n}_\perp$ denotes the unit vector on Γ. $\vec{J}_{n_0}^{\sim}$ and $\vec{J}_{p_0}^{\sim}$ are the zeroth-order terms of the smooth parts of the (scaled) electron and hole current densities.

$$\begin{aligned} J_{n_0}^{\sim} &= \text{grad}\, n_0^{\sim} - n_0^{\sim} \cdot \text{grad}\, \Psi_0^{\sim} \\ J_{p_0}^{\sim} &= \text{grad}\, p_0^{\sim} + p_0^{\sim} \cdot \text{grad}\, \Psi_0^{\sim}. \end{aligned} \tag{2.29}$$

Equations (2.22)-(2.24) together with (2.25)-(2.28) and the boundary conditions (2.6)-(2.9) define the reduced problem whose solution is an $0(\lambda)$ approximation to the full solution away from Γ. As we will see in the next section, the reduced problem is a useful tool for the development and analysis of numerical methods since it (especially the conditions (2.25)-(2.28)) has to be solved implicitly by any discretization method which requires a reasonable number of grid points.

The equations for the rapidly varying parts $w_i^{\wedge}$ reduce to ordinary differential equations. This means that only derivatives with respect to the "fast" variable t/λ occur. Since the rate of decay of $w_i^{\wedge}$ depends heavily on Ψ, the width of the layer grows with the applied voltage; a fact which is absolutely well known by device physicists, but which becomes nicely apparent by the singular perturbation approach.

III. Numerical Solution of the Semiconductor Equations

In this section we discuss some of the problems occurring in the numerical solution of the semiconductor equations and the analysis of existing numerical methods. From the viewpoint of numerical analysis, there are essentially four major topics to be considered. The first one is the type of discretization to be used. There exist programs for both Finite Element and Finite Difference discretizations of the system (2.1)-(2.5). As outlined in the previous chapter the solution exhibits a smooth behavior in some subregions of the domain whereas in others it varies rapidly. Thus a nonuniform mesh is mandatory and adaptive mesh refinement is desirable. So the second topic is the question how to set up the mesh refinement algorithm, i.e., which quantities have to be used to control the mesh. Each type of discretization will lead to a large sparse system of nonlinear equations and so the solution of this system is the third topic. As the fourth topic, we discuss the linear equation solvers which have to be used in topic three. For topics one to three, many methods have been designed especially for the semiconductor equations. These points will be discussed in this section. For topic four, standard numerical analysis is commonly used and so its discussion will be deferred to Section IV. For the sake of simplicity in nomenclature we shall only consider the two-dimensional case in this chapter. However, all results given in the following can be generalized to three dimensions in a straightforward manner. So, the equations are posed in a domain D of $\mathbb{R}^2$ and $\vec{x} = (x, y)^T$ denotes the independent variable.

3.1. Discretization Schemes

Using Finite Elements or Finite Differences, one has to take into account that Poisson's equation (2.1) is of a different type than the continuity equations. Poisson's equation—in the scaling of Markowich [40]—using the variables (Ψ, u, v)

$$\lambda^2 \cdot \text{div grad}\, \Psi = e^{\Psi} \cdot u - e^{-\Psi} \cdot v - C \tag{3.1}$$

is a singularly perturbed elliptic problem whose right-hand side has a positive derivative with respect to Ψ. Thus it is of a standard form (as discussed in, e.g., [22]) except for the discontinous or exponentially graded term C. Equations of that type are generally well behaved and it suffices to apply a usual discretization scheme. In the case of Finite Differences equation (3.1) is discretized by

$$\lambda^2 \cdot (\text{div grad}\, \Psi)_{ij} = n_{ij} - p_{ij} - C(x_i, y_j) \tag{3.2}$$

$$\begin{aligned} E^x_{i+1/2,j} &= (\Psi_{i+1,j} - \Psi_{i,j})/h_i \\ E^y_{i,j+1/2} &= (\Psi_{i,j+1} - \Psi_{i,j})/k_j \\ h_i &= x_{i+1} - x_i \\ k_j &= y_{j+1} - y_j \end{aligned} \tag{3.3}$$

$$\begin{aligned} (\text{div grad}\, \Psi)_{i,j} = {} & 2 \cdot (E^x_{i+1/2,j} - E^x_{i-1/2,j})/(h_i + h_{i-1}) \\ & + 2 \cdot (E^y_{i,j+1/2} - E^y_{i,j-1/2})/(k_j + k_{j-1}). \end{aligned} \tag{3.4}$$

Here Ψ_{ij}, n_{ij}, and p_{ij} denote the approximations to Ψ, n, and p at the gridpoint (x_i, y_j). $E^x_{i+1/2,j}$ denotes the value of $\partial\Psi/\partial x$ at $(x_{i+1/2} = (x_i + x_{i+1})/2,\ y_j)$. $E^y_{i,j+1/2}$ denotes the value of

$\partial\Psi/\partial y$ at $(x_i, y_{j+1/2} = (y_j + y_{j+1})/2)$. If one of the neighboring gridpoints (x_{i+1}, y_j), (x_{i-1}, y_j), (x_i, y_{j+1}), (x_i, y_{j-1}) does not exist—as possible in a terminating line approach [1], [2] or in the Finite Boxes approach [24]–(3.4) has to be modified. We will go into some detail concerning these modifications in the next section. In the case of Finite Elements, classical shape functions can be used (i.e., linear shape functions for triangular elements, bilinear shape functions for rectangular elements).

It turns out that the discretization of the continuity equations is more crucial than the discretization of Poisson's equation. The usual error analysis of discretization methods provides an error estimate of the form

$$\max |w_h - w| <= c \cdot H \tag{3.5}$$

where w_h denotes the numerical approximation to $w(x, y) = (\Psi, n, p)^T$. H denotes the maximal gridspacing. The constant c will, in general, depend on the higher order derivatives of w. The singular perturbation analysis [41] shows that derivatives of $\hat{\Psi}$, $\hat{n}$ and $\hat{p}$ in (2.21) are of magnitude $0(\lambda^{-3})$-$0(\lambda^{-4})$ locally near the junction (λ is defined in (2.17)). Reference [41] shows also that, even if a nonuniform mesh is used, the amount of gridpoints required to equidistribute the error term in (3.5) can be proportional to λ^{-2} which is of course prohibitive. Therefore, a discretization scheme is needed where the constant c in (3.5) does not depend on the higher derivatives of the rapidly varying terms $\hat{\Psi}$, $\hat{n}$, and $\hat{p}$. For the case of Finite Differences such a scheme was given by Scharfetter and Gummel [50]. They approximate

$$\vec{J}_n = \text{grad}\, n - n \cdot \text{grad}\, \Psi \tag{3.6}$$

$$\text{div}\, \vec{J}_n = \partial J_n^x/\partial x + \partial J_n^y/\partial y = R \tag{3.7}$$

by

$$\begin{aligned} J^x_{n_{i+1/2,j}} &= \zeta((\Psi_{i+1,j} - \Psi_{i,j})/2) \cdot (n_{i+1,j} - n_{i,j})/h_i \\ &\quad - (n_{i,j} + n_{i+1,j})/2 \cdot (\Psi_{i+1,j} - \Psi_{i,j})/h_i \\ J^y_{n_{i,j+1/2}} &= \zeta((\Psi_{i,j+1} - \Psi_{i,j})/2) \cdot (n_{i,j+1} - n_{i,j})/k_j \\ &\quad - (n_{i,j} + n_{i,j+1})/2 \cdot (\Psi_{i,j+1} - \Psi_{i,j})/k_j \end{aligned} \tag{3.8}$$

$$\zeta(s) = s \cdot \coth(s)$$

$$\begin{aligned} R_{i,j} &= 2 \cdot (J^x_{n_{i+1/2,j}} - J^x_{n_{i+1/2,j}})/(h_i + h_{i-1}) \\ &\quad + 2 \cdot (J^y_{n_{i,j+1/2}} - J^y_{n_{i,j-1/2}})/(k_j + k_{j-1}). \end{aligned} \tag{3.9}$$

$J^x_{n_{i+1/2,j}}$ denotes the value of J_n^x at $(x_{i+1/2} = (x_i + x_{i+1})/2, y_j)$. $J^y_{n_{i,j+1/2}}$ denotes the value of J_n^y at $(x_i, y_{j+1/2} = (y_j + y_{j+1})/2)$. The continuity equation for holes is discretized analogously. Scharfetter and Gummel give a physical reasoning for the derivation of their scheme. Markowich *et al.*, [41] proved that in one dimension the Scharfetter-Gummel scheme is uniformly convergent. That means that the error constant c in (3.5) does not depend on the derivatives of $\hat{\Psi}$, $\hat{n}$, and $\hat{p}$ in (2.21) and, therefore, not on λ. For two dimensions, [41] shows that the choice $\zeta(s) = s \cdot \coth(s)$ is necessary for uniform convergence. Exponentially fitted schemes like the Scharfetter-Gummel scheme have been analyzed by Kellog [33], [34] and Doolan [17] (for different classes of problems). The reason for the uniform convergence of these schemes is that inside the p-n-junction layers, the interface conditions (2.25) and (2.26) are satisfied automatically if $|\text{grad}\, \Psi|$ is large and the gridspacing is not $0(\lambda)$.

The results for Finite Difference schemes suggest that a similar approach (like the exponentially fitted schemes) should be used in the case of Finite Elements. This fact has been intuitively observed by Engel [21] for the one-dimensional case. A modeling group at IBM has tried to make use of the Scharfetter-Gummel scheme for Finite Elements in two and three space dimensions [8], [9], [12]. However, we have the impression that their approach still needs quite a bit of analysis, although it has also been used effectively by other modelists e.g., [49]. Macheck [36] has tried to develop a more rigorous discretization for Finite Elements using exponentially fitted shape functions. He uses classical bilinear shape functions for Ψ and

$$\begin{aligned} \alpha_1(x,y) &= [1 - \varphi_1(x,y)] \cdot [1 - \varphi_2(x,y)] \\ \alpha_2(x,y) &= \varphi_1(x,y) \cdot [1 - \varphi_2(x,y)] \\ \alpha_3(x,y) &= \varphi_1(x,y) \cdot \varphi_2(x,y) \\ \alpha_4(x,y) &= [1 - \varphi_1(x,y)] \cdot \varphi_2(x,y) \end{aligned} \tag{3.11}$$

for n, and

$$\begin{aligned} \rho_1(x,y) &= [1 - \sigma_1(x,y)] \cdot [1 - \sigma_2(x,y)] \\ \rho_2(x,y) &= \sigma_1(x,y) \cdot [1 - \sigma_2(x,y)] \\ \rho_3(x,y) &= \sigma_1(x,y) \cdot \sigma_2(x,y) \\ \rho_4(x,y) &= [1 - \sigma_1(x,y)] \cdot \sigma_2(x,y) \end{aligned} \tag{3.12}$$

for p where

$$\begin{aligned} \varphi_1(x,y) &= f\left(x, \frac{\partial\Psi}{\partial x}\right) \\ \varphi_2(x,y) &= f\left(y, \frac{\partial\Psi}{\partial y}\right) \\ \sigma_1(x,y) &= f\left(x, -\frac{\partial\Psi}{\partial x}\right) \\ \sigma_2(x,y) &= f\left(y, -\frac{\partial\Psi}{\partial y}\right) \end{aligned} \tag{3.13}$$

with

$$f(x,a) = (\exp(ax) - 1)/(\exp(a) - 1). \tag{3.14}$$

The advantage of these shape functions is that they nicely accomodate the layer behavior of the solution. They degenerate into the ordinary bilinear shape functions when the electric potential is constant. In order to be able to switch from coarse to fine grid spacing in different subdomains transition elements have to be used (as outlined in the next subsection). However, no theoretical investigations have been carried out so far to analyze the uniform convergence properties of this method.

3.2. Grid Construction

Since subregions of strong variation of Ψ, n, and p alternate with regions where these quantities behave smoothly (i.e., their

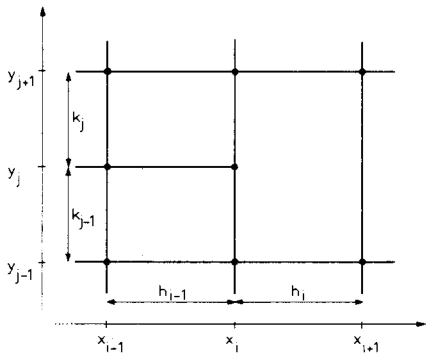


Fig. 2. A typical Finite Boxes configuration.

gradients are small) different mesh sizes are mandatory in these subregions. Thus the discretization scheme should be able to switch locally from a coarser to a finer grid. For the exponentially fitted (Scharfetter-Gummel) Finite Difference discretization schemes this can be done by the Finite Boxes approach [24]. Grid lines can terminate when the mesh is likely to be made more coarse (cf. (Fig. 2). The point (x_{i+1}, y_j) does not belong to the mesh. Thus the equations for the point (x_i, y_j) have to be modified since $\Psi_{i+1,j}$, $n_{i+1,j}$, and $p_{i+1,j}$ are not available. This is done by proper interpolation between the $(j-1)$st and $(j+1)$st y level. So $(\text{div grad } \Psi)_{ij}$ is approximated by

$$\begin{aligned}(\text{div grad } \Psi)_{i,j} = 2 \cdot ((&k_{j-1} \cdot E^x_{i+1/2,j+1} + k_j \\ &\cdot E^x_{i+1/2,j-1})/(k_j + k_{j-1}) \\ &- E^x_{i-1/2,j})/(h_i + h_{i-1}) \\ &+ 2 \cdot (E^y_{i,j+1/2} - E^y_{i,j-1/2})/(k_j + k_{j-1}).\end{aligned} \tag{3.15}$$

$E^x_{i-1/2,j}$, $E^y_{i,j+1/2}$, etc., are defined in (3.3). The continuity equations are approximated by

$$\begin{aligned}2 \cdot ((&k_{j-1} \cdot J^x_{n_{i+1/2,j+1}} + k_j \cdot J^x_{n_{i+1/2,j-1}})/(k_j + k_{j-1}) \\ &- J^x_{n_{i-1/2,j}})/(h_i + h_{i-1}) + 2 \cdot (J^y_{n_{i,j+1/2}} \\ &- J^y_{n_{i,j-1/2}})/(k_j + k_{j-1}) = R_{i,j}.\end{aligned} \tag{3.16}$$

$J^x_{n_{i-1/2,j}}$, $J^y_{n_{i,j+1/2}}$, etc., are defined in (3.8). For reasons of numerical stability, only one gridline is allowed to terminate at a box. This approach is a generalization of the "Terminating Line" approach introduced by Adler [1], [2] as already mentioned.

In the Finite Element approach of Macheck [36] transition elements composed of three triangles are used to increase mesh coarseness locally (cf. Fig. 3). Within these triangles a different set of shape functions has to be used. They are derived by holding the current densities $\vec{J}_n$ and $\vec{J}_p$ constant along the edges of a triangle similar to the approach of [10].

In the Finite Element, as well as in the Finite Difference (Boxes) approach, the question arises which criteria should be used to generate the mesh. If the user of a simulation program has to define his elements or nodes *a priori* as input parameters, this could perhaps be done by experience [10]. However, if—as it is the case for modern user-oriented programs—an adaptive mesh selection is desired, mathematically formulated criteria are a *sine qua non*. Generally such criteria should satisfy two conditions. Firstly, they should not cause the program to construct more gridpoints/elements than necessary to achieve a certain accuracy. Secondly, they should guarantee that a prescribed relative accuracy δ is really achieved once they are satisfied. A usual way to design adaptive mesh refinement procedures is to equidistribute the local truncation error of the discretization scheme. In the case of Finite Differences, this error is proportional to the mesh size and the third and fourth derivatives of Ψ, n, and p. Markowich [41], however, showed that it is practically not possible to equidistribute this quantity. In the case of a simple MOS transistor $0(\delta^{-2}\lambda^{-2})$ gridpoints would be required. On the other hand, the singular perturbation analysis shows that the solution of the difference scheme approximates the solution of the reduced problem (2.22)-(2.24) even if this criterion is not satisfied inside the layer regions (inversion layer and space charge regions). Therefore, the quantity to be equidistributed is the discretization error of Poisson's equation (i.e., the partial derivatives of the space charge times the mesh sizes). This equidistribution can be relaxed inside the p-n-junction layers by, e.g., simply limiting the number of gridpoints there.

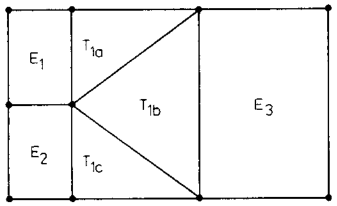


Fig. 3. A transition element to coarsen a mesh.

3.3. Linearization Schemes

Each discretization scheme (Finite Differences or Finite Elements) will lead to a large sparse system of nonlinear equations to be solved. The theory of iterative methods to solve these equations is to a large extent independent of the discretization used, and so it is convenient to view the whole problem as solving a nonlinear system of equations iteratively by solving linear systems. The existing numerical methods can essentially be divided into two classes. The first approach, a block nonlinear iteration algorithm, is due to Gummel [29] and uses the fact that the current relations are linear in the variables u and v (as defined in (2.11)). In these variables the equations become (again we use the scaling of [36])

$$\lambda^2 \cdot \text{div grad } \Psi = e^{\Psi} \cdot u - e^{-\Psi} \cdot v - C \tag{3.17}$$

$$\text{div } \vec{J}_n = R, \qquad \vec{J}_n = e^{\Psi} \cdot \text{grad } u \tag{3.18}$$

$$\text{div } \vec{J}_p = -R, \qquad \vec{J}_p = -e^{-\Psi} \cdot \text{grad } v. \tag{3.19}$$

Gummel's approach works as follows. Given $(\Psi, u, v)^k$, Ψ^{k+1} is computed by solving

$$\lambda^2 \cdot \text{div grad } \Psi^{k+1} = e^{\Psi^{k+1}} \cdot u^k - e^{-\Psi^{k+1}} \cdot v^k - C \tag{3.20}$$

subject to the appropriate boundary conditions. Then u^{k+1} and v^{k+1} are computed from

$$\begin{aligned} \operatorname{div} \vec{J}_n^{k+1} &= R(\operatorname{grad} \Psi^{k+1}, u^k, v^k), \\ \vec{J}_n^{k+1} &= e^{\Psi^{k+1}} \cdot \operatorname{grad} u^{k+1} \end{aligned} \tag{3.21}$$

$$\begin{aligned} \operatorname{div} \vec{J}_p^{k+1} &= -R(\operatorname{grad} \Psi^{k+1}, u^k, v^k), \\ \vec{J}_p^{k+1} &= -e^{-\Psi^{k+1}} \cdot \operatorname{grad} v^{k+1} \end{aligned} \tag{3.22}$$

together with the boundary conditions for u and v. Equations (3.21) and (3.22) are two decoupled linear equations for u^{k+1} and v^{k+1}. Poisson's equation (3.2) is nonlinear in this setting and therefore it has to be solved iteratively itself in each step by a Newton-like method. Since Newton's method is an inner iteration within the overall iteration process (3.20)-(3.22), it may not be necessary to let this inner iteration "fully converge" [27]. It could for instance be considered necessary to do only one Newton step for each iteration. This would lead to the linear equation

$$\begin{aligned} \lambda^2 \cdot \operatorname{div} \operatorname{grad} \Psi^{k+1} &= (e^{\Psi^k} \cdot u^k + e^{-\Psi^k} \cdot v^k) \cdot (\Psi^{k+1} - \Psi^k) \\ &\quad + e^{\Psi^k} \cdot u^k - e^{-\Psi^k} \cdot v^k - C \end{aligned} \tag{3.23}$$

instead of (3.20). The advantage of Gummel's method is obvious. (3.20)-(3.22) can be solved sequentially, which decreases the required amount of storage and computing time drastically for each step. However, bad convergence properties can be observed in the case of high currents. This is explained by viewing (3.20)-(3.22) as iterating the map M: $(u^k, v^k) \to (u^{k+1}, v^{k+1})$ where the evaluation of M involves the solution of (3.20). Then the norm of the linearization of M (as an operator acting in the appropriate spaces) at the fixpoint $M(u^*, v^*) = (u^*, v^*)$ is proportional to the current densities [42].

The second approach to the solution of the nonlinear equations (2.1)-(2.5) is a damped modified Newton method. To solve the geneal equation $F(x) = 0$, one computes the sequence $\langle x^k \rangle$ by

$$M^k \cdot \delta^k = -F(x^k), \qquad x^{k+1} = x^k + t^k \cdot \delta^k. \tag{3.24}$$

For the usual Newton method, $M^k = F'(x^k)$ and $t^k = 1$ holds. Bank and Rose [4] have given criteria for the choice of the damping parameters t^k which guarantee global convergence. Moreover they investigate how well δ^k has to approximate the classical Newton step in order to get a certain rate of convergence. They find that the rate of convergence is p $(1 < p < 2)$ if

$$|M^k \cdot \delta^k + F(x^k)| = 0(|F(x^k)|^p) \tag{3.25}$$

holds asymptotically for $k \to \infty$. Alternatively, Bank and Rose [3] suggested $M^k = \lambda^k I + F'(x^k)$ where λ^k is proportional to $|F(x^k)|$. Franz [24] tested this method with good success. However, he additionally chose damping parameters t^k according to Deuflhard [15], [16].

Since this approach has the disadvantage that all three equations are solved simultaneously—and, therefore, the storage requirements are fairly large—we suggest a Block-Newton-SOR (successive overrelaxation) method [24]. Defining $F = (F_1, F_2, F_3)^T$, Newton's method at step k is

$$\left|\begin{matrix} \dfrac{\partial F_1}{\partial \Psi} & \dfrac{\partial F_1}{\partial n} & \dfrac{\partial F_1}{\partial p} \\ \dfrac{\partial F_2}{\partial \Psi} & \dfrac{\partial F_2}{\partial n} & \dfrac{\partial F_2}{\partial p} \\ \dfrac{\partial F_3}{\partial \Psi} & \dfrac{\partial F_3}{\partial n} & \dfrac{\partial F_3}{\partial p} \end{matrix}\right|^k \cdot \left|\begin{matrix} \delta\Psi^k \\ \delta n^k \\ \delta p^k \end{matrix}\right| = -\left|\begin{matrix} F_1(\Psi^k, n^k, p^k) \\ F_2(\Psi^k, n^k, p^k) \\ F_3(\Psi^k, n^k, p^k) \end{matrix}\right|. \tag{3.26}$$

Under the assumption that the Jacobian is definite, one can use a classical block iteration scheme (iteration index m) for the solution of the kth Newton step

$$\begin{aligned} &\left|\begin{matrix} \dfrac{\partial F_1}{\partial \Psi} & 0 & 0 \\ \dfrac{\partial F_2}{\partial \Psi} & \dfrac{\partial F_2}{\partial n} & 0 \\ \dfrac{\partial F_3}{\partial \Psi} & \dfrac{\partial F_3}{\partial n} & \dfrac{\partial F_3}{\partial p} \end{matrix}\right|^k \cdot \left|\begin{matrix} \delta\Psi^k \\ \delta n^k \\ \delta p^k \end{matrix}\right|^{m+1} \\ &= -\left|\begin{matrix} F_1(\Psi^k, n^k, p^k) \\ F_2(\Psi^k, n^k, p^k) \\ F_3(\Psi^k, n^k, p^k) \end{matrix}\right| - \left|\begin{matrix} 0 & \dfrac{\partial F_1}{\partial n} & \dfrac{\partial F_1}{\partial p} \\ 0 & 0 & \dfrac{\partial F_2}{\partial p} \\ 0 & 0 & 0 \end{matrix}\right|^k \left|\begin{matrix} \delta\Psi^k \\ \delta n^k \\ \delta p^k \end{matrix}\right|^m. \end{aligned} \tag{3.27}$$

Since the coefficient matrix of (3.27) is block lower triangular, one can decouple the elimination process into three linear systems, (3.28)-(3.30), which have to be solved sequentially

$$\begin{aligned} \frac{\partial F_1^k}{\partial \Psi} \cdot \delta\Psi^{km+1} &= -F_1(\Psi^k, n^k, p^k) \\ &\quad - \frac{\partial F_1^k}{\partial n} \cdot \delta n^{km} - \frac{\partial F_1^k}{\partial p} \cdot \delta p^{km} \end{aligned} \tag{3.28}$$

$$\begin{aligned} \frac{\partial F_2^k}{\partial n} \cdot \delta n^{km+1} &= -F_2(\Psi^k, n^k, p^k) \\ &\quad - \frac{\partial F_2^k}{\partial \Psi} \cdot \delta\Psi^{km+1} - \frac{\partial F_2^k}{\partial p} \cdot \delta p^{km} \end{aligned} \tag{3.29}$$

$$\begin{aligned} \frac{\partial F_3^k}{\partial p} \cdot \delta p^{km+1} &= -F_3(\Psi^k, n^k, p^k) \\ &\quad - \frac{\partial F_3^k}{\partial \Psi} \cdot \delta\Psi^{km+1} - \frac{\partial F_3^k}{\partial n} \cdot \delta n^{km+1}. \end{aligned} \tag{3.30}$$

This iteration method has (like Gummel's method) the advantage that the equations can be solved sequentially. To end up with the Block-Newton-SOR method, one has to resubstitute the series expansions on the right-hand side of (3.28)-

(3.30) and to introduce a relaxation parameter ω:

$$\frac{\partial F_1^k}{\partial \Psi} \cdot \delta\Psi^{km+1} = -\omega \cdot F_1(\Psi^k, n^k + \delta n^{km}, p^k + \delta p^{km}) \tag{3.31}$$

$$\frac{\partial F_2^k}{\partial n} \cdot \delta n^{km+1} = -\omega \cdot F_2(\Psi^k + \delta\Psi^{km+1}, n^k, p^k + \delta p^{km}) \tag{3.32}$$

$$\frac{\partial F_3^k}{\partial p} \cdot \delta p^{km+1} = -\omega \cdot F_3(\Psi^k + \delta\Psi^{km+1}, n^k + \delta n^{km+1}, pk). \tag{3.33}$$

This method converges linearly [48]. However, we still have to perform through investigations in order to properly judge the convergence properties.

IV. Solution of Linear Systems

For any of the linearization procedures which have been outlined in the last chapter, a large sparse linear equation system (4.1) has to be solved repeatedly

$$A \cdot x = b. \tag{4.1}$$

A has been derived by linearizing discretized PDE's. Hence A has only five to nine nonzero entries per row and block (the blocks are defined in (3.26); A is very sparse. For the solution of these special types of linear systems of equations, two classes of methods can, in principle, be used–direct methods which are based on elimination and iterative methods. An excellent survey on that subject has been published recently by Duff [18]. Classical Gaussian elimination is not feasible for our systems of equations because the rank of A in (4.1) is very large and A has many coefficients which are zero. Therefore, modifications of the classical Gaussian elimination algorithm have to be introduced to account for the zero entries. There exist quite a few activities on that subject (c.f., [19]) and powerful algorithms which treat the nonzero coefficients only are available (the so-called sparse matrix codes). Another serious drawback of direct methods lies in the fact that the upper triangular matrix which is created by the elimination process has to be stored for back substitution. This matrix has usually more nonzero entries than the matrix A. Therefore, memory requirement of direct methods is substantial. One advantage of the linear systems obtained from the discretised semiconductor equations is that no pivoting in order to maintain numerical stability is needed. In spite of all the drawbacks of direct methods, their major advantage is high accuracy of the solution. However, we feel that for the semiconductor problems iterative algorithms should be emphasized. Nevertheless, we and many others have observed difficulties with respect to the convergence speed of iterative methods, so that the direct methods, which require an exactly predictable amount of computer resources, will always stay in consideration.

The fundamental idea of relaxation methods (which are the best established iterative methods) is the splitting of the coefficient matrix A (4.1) into three matrices D, E, F (4.2)

$$A = D - E - F \tag{4.2}$$

where D denotes the diagonal entires of A, $-E$ denotes a lower triangular matrix which consists of all subdiagonal entries of A, and $-F$ denotes an upper triangular matrix which consists of all superdiagonal entries of A.

With an arbitrary nonsingular matrix B which has the same rank as A, the linear system (4.1) can be rewritten as

$$B \cdot x + (A - B) \cdot x = b. \tag{4.3}$$

One obtains an iterative scheme by setting

$$B \cdot x^{k+1} = b - (A - B) \cdot x^k. \tag{4.4}$$

Equation (4.4) can be solved for x^{k+1}

$$x^{k+1} = (I - B^{-1} \cdot A) \cdot x^k + B^{-1} \cdot b. \tag{4.5}$$

The scheme (4.5) will converge if condition (4.6) holds

$$\rho(I - B^{-1} \cdot A) < 1. \tag{4.6}$$

Equation (4.6) is a necessary and sufficient condition where ρ denotes the spectral radius [64]. Any relaxation method can be derived by differently choosing the matrix B from the splitting of A (4.2). The simplest scheme, the point-Jacobi method, uses D for B. Matrix D is a diagonal matrix and, therefore, is easily invertible. The Gauss-Seidel method uses $D - E$ for B. The matrix $D - E$ is a lower triangular matrix. Therefore, one has only to perform a forward substitution process for its inversion. The SOR uses a parameter ω within the range $]0, 2[$. The iteration matrix B is defined

$$B = D/\omega - E. \tag{4.7}$$

Since B is again a lower triangular matrix, its inversion is instantly reduced to a substitution.

The major advantage of these iterative methods lies in their simplicity. They are very easy to program and demand only small amounts of memory. As already noted, they converge if condition (4.6) holds. However, this is generally difficult to prove. A sufficient condition for convergence is that A is positive definite (4.8), which is the normal case for five-point-star discretized PDE's.

$$x^T \cdot A \cdot x > 0 \qquad \text{for all} \quad x \neq 0. \tag{4.8}$$

It should be noted again here that the current relations and continuity equations are not self-adjoint if (Ψ, n, p) are used as variables (see (2.10), (2.11)). However, the transformation

$$n = e^{\Psi} \cdot u, \qquad p = e^{-\Psi} \cdot v \tag{4.9}$$

results in a similarity transformation of the iteration matrix in (4.6). Thus the spectral radius of the iteration matrix is not influenced and the same convergence properties are obtained as if the system had been discretized in its self-adjoint form with (Ψ, u, v) as variables.

Some point-iterative schemes can by accelerated quite remarkably with the conjugate gradient method or the Chebyshev method. An excellent survey on these topics can be found in [28].

Various activities can be observed for the development of more powerful algorithms with the advantages of iterative schemes. One of the best known algorithms which has been established in semiconductor device analysis is Stone's strongly implicit procedure [58]. Stone's idea was to modify the original coefficient matrix A by adding a matrix N (whose norm is much smaller than the norm of A) so that a factorization of $(A + N)$ involves less computational effort than the standard decomposition of A. Assuming this has been done, the development of an iterative procedure is then fairly straightforward because the equation can be written as

$$(A + N) \cdot x = (A + N) \cdot x + (b - A \cdot x) \tag{4.10}$$

which suggests the iterative procedure

$$(A + N) \cdot x^{k+1} = (A + N) \cdot x^k + (b - A \cdot x^k). \tag{4.11}$$

When the right-hand side is known and if $(A + N)$ can be factorized easily, (4.11) gives an efficient method for directly solving for x^{k+1}. Furthermore, one would intuitively expect a rapid rate of convergence if N is sufficiently small compared to A. We will refrain from explaining in detail Stone's suggestion of how to choose the perturbation matrix N because this has been done thoroughly in many publications e.g., [23], [55], [58]. A major disadvantage of Stone's method is that it is only applicable for linear systems obtained by a classical Finite Difference discretization. It is not applicable for systems obtained by the Finite Boxes approach or the general Finite Element approach.

There exist a few algorithms which are similar to Stone's method in terms of underlying ideas. The most attractive are the method of Dupont *et al.*, [20], the "alternating direction implicit" methods, e.g., [6], [23], [66], and the Fourier methods [57], [64]. However, more of these sophisticated algorithms lack general applicability.

No matter which iterative method is used, one has to deal with the question of an appropriate termination (convergence) criterion. Usually (4.12) is applied with a properly chosen relative accuracy ϵ

$$\left|x^{k+1} - x^k\right| < \epsilon \cdot \left|x^{k+1}\right|. \tag{4.12}$$

Since increments still accumulate when (4.12) is already satisfied, we suggest using (4.13) instead of (4.12)

$$\left|x^{k+1} - x^k\right| < \epsilon \cdot \left|x^{k+1}\right| \cdot (1 - \rho(G)). \tag{4.13}$$

$\rho(G)$ can be estimated as

$$\lim_{k \to \infty} \left|x^{k+1} - x^k\right| / \left|x^k - x^{k-1}\right|.$$

One disadvantage of all strongly implicit methods and also the direct methods is that they cannot be implemented efficiently on a computer with a pipeline architecture (vector processor). Some comments on that subject have been given in [18].

V. A Glimpse on Results

As an illustrative example, a relatively simple structure—a two-dimensional diode—is chosen. Fig. 4 shows the doping profile as birds-eye-view plot. A substrate with 10^{14} cm^{-3} acceptor concentration and an exponentially graded n-region with 10^{19} cm^{-3} maximum doping is assumed. The initial mesh is automatically generated from the doping profile and the geometry definition. The simulation domain (device geometry) is a square of 100 X 100 μm size. At the n-region, an ohmic contact with length 20 μm is assumed. The substrate is fully contacted. The initial mesh for a Finite Boxes program is shown in Fig. 5 and for a Finite Element program in Fig. 6. The point allocation is identical for both representations. The grid consists of 121 points versus 178 when all gridlines are extended throughout the device. This clearly demonstrates the advantage of the Finite Boxes approach. In Finite Element representation one has to deal with 80 rectangular elements and 17 transition elements which consist of 51 triangles.

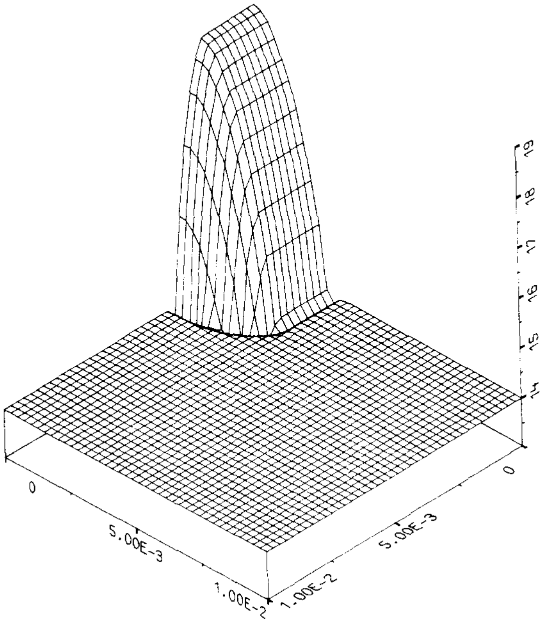


Fig. 4. Doping profile [cm^{-3}] (log).

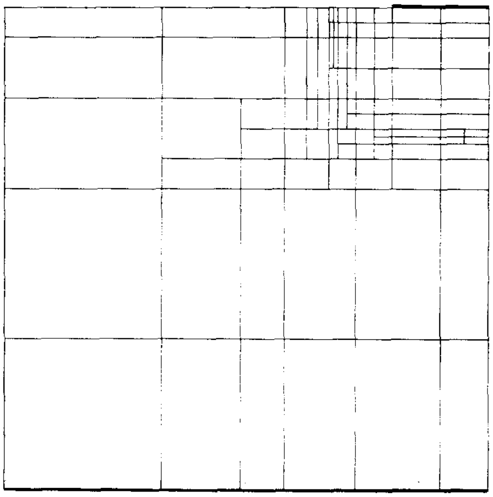


Fig. 5. Initial mesh in Finite Boxes interpretation.

Fig. 7 shows the final grid for an operating condition of 0.7-V forward bias in Finite Boxes representation. This mesh is obtained after several adaption processes using the criteria given in Section III. It consists of 270 points (versus 480 for the classical approach). In Fig. 8 the potential distribution is

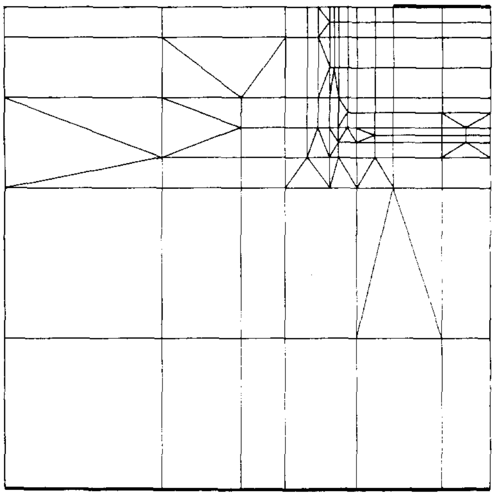


Fig. 6. Initial mesh in Finite Element interpretation.

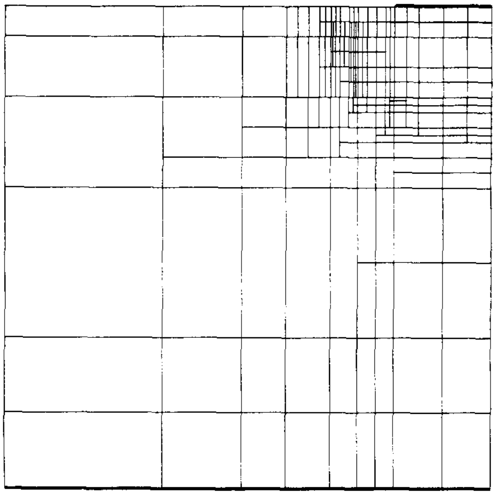


Fig. 7. Final mesh for 0.7-V forward bias (Finite Boxes).

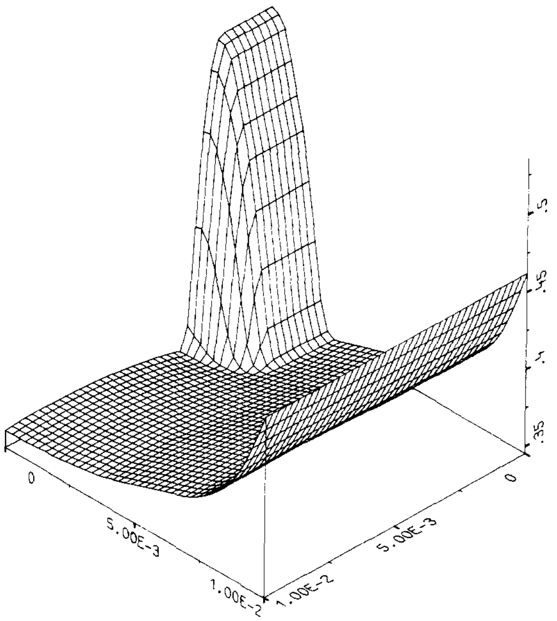


Fig. 8. Potential distribution (0.7 V) [V] (lin).

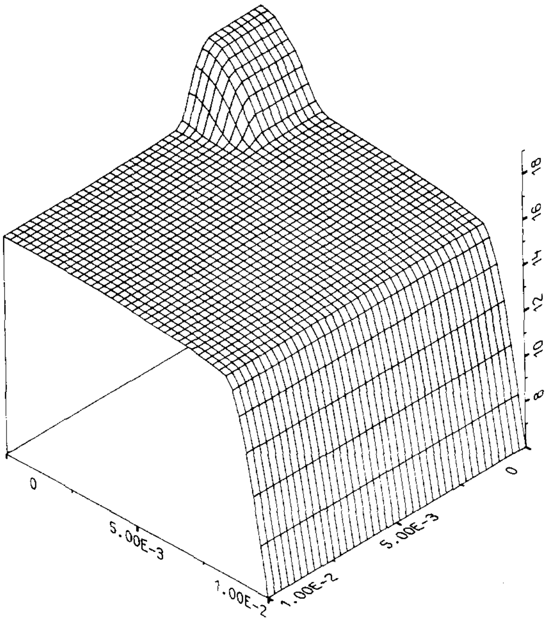


Fig. 9. Electron concentration (0.7 V) [cm^{-3}] (log).

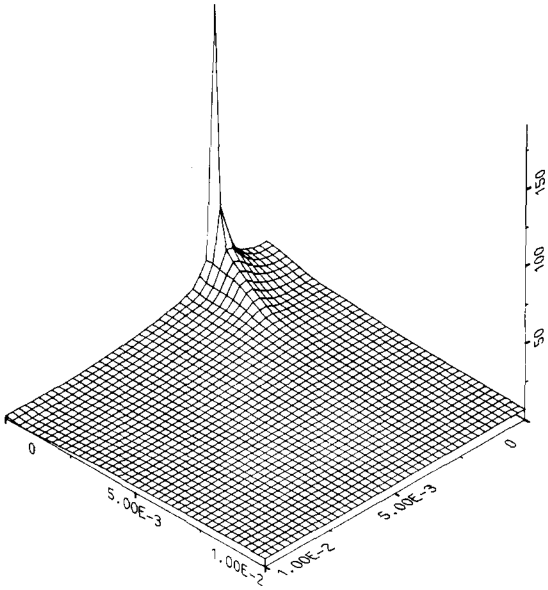


Fig. 10. Electron current density (0.7 V) [$\mathrm{A/cm}^2$] (lin).

drawn. From this plot, and even better from the electron density (Fig. 9), one can nicely deduce the effects of high injection e.g., the substrate is flooded with carriers. Fig. 10 shows the magnitude of the electron current density. The peak value is about 180 $\mathrm{A/cm}^2$. The sharply pronounced peak which exists at the transition of the Dirichlet boundary condition to the Neumann boundary condition corresponds to a singularity of the carrier densities. Physically interpreted, this effect is well known as contact-corner-current-crowding.

Fig. 11 shows the final grid for an operating condition of -20-V (reverse) bias in Finite Element representation. This mesh consists of 363 points (625 for classical Finite Differences) which correspond to 277 rectangular elements and 41 transition elements (123 triangles). The electron density for this operating point is given in Fig. 12. One nicely observes the depletion region and the typical shape of the drop of the elec-

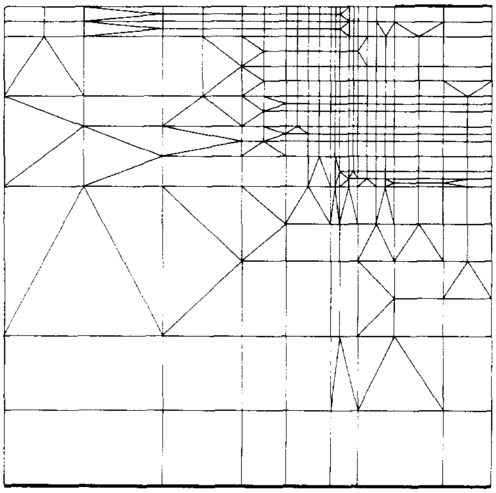


Fig. 11. Final mesh for 20-V reverse bias (Finite Elements).

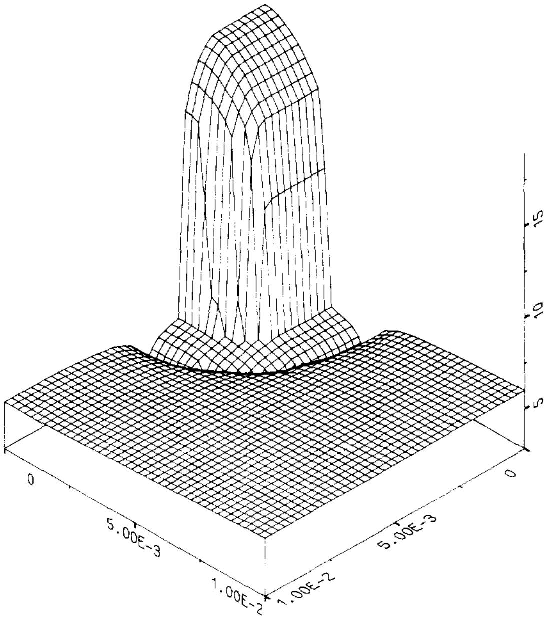


Fig. 12. Electron concentration (-20 V) [cm^{-3}] (log).

tron density in that region owing to thermal generation. In Fig. 13 the magnitude of the electron current density is drawn. The singularity at the contact corner is, although still existent, not so pronounced. Note that there are about seven orders of magnitude difference in the peak value compared to Fig. 10.

VI. Conclusion

In this paper we have presented an analysis of the steady-state semiconductor equations and the impact of this analysis on the design of device simulation programs. By appropriate scaling we have transformed the semiconductor equations into a singularly perturbed elliptic system with nonsmooth data. Information obtained from the singular perturbation analysis has been used to investigate stability and convergence of discretization schemes with particular emphasis on the adaptive construction of efficient grids. We have reviewed algorithms for the solution of nonlinear and linear systems of the discretized semiconductor equations. An example has demonstrated the power and flexibility a device simulation program can achieve when using the information we have presented for program design.

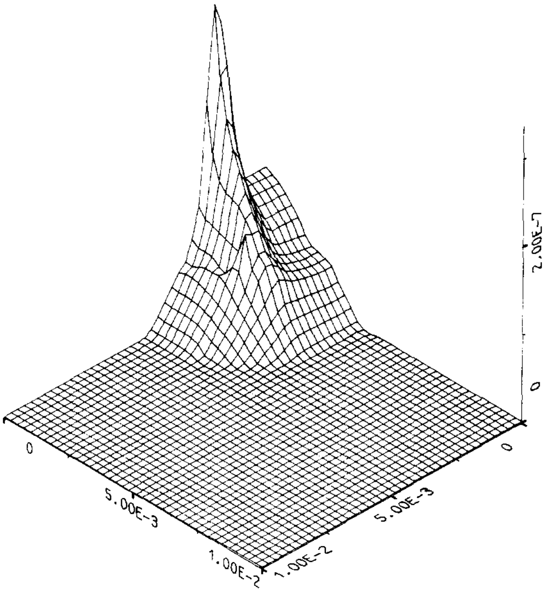


Fig. 13. Electron current density (-20 V) [A/cm^2] (lin).

Acknowledgment

Helpful discussions with Dr. P. Markowich and Prof. Dr. H. Pötzl, who also critically read the manuscript, were highly appreciated. We should also like to thank Dipl.Ing.A. Franz and Dipl.Ing.G. Franz for providing the computer plots.

References

[1] M. S. Adler, "A method for achieving and choosing variable density grids in finite difference formulations and the importance of degeneracy and band gap narrowing in device modeling," in *Proc. NASECODE I Conf.*, 1979, pp. 3-30.

[2] ——, "A method for terminating mesh lines in finite difference formulations of the semiconductor device equations," *Solid-State Electron.*, vol. 23, pp. 845-853, 1980.

[3] R. E. Bank and D. J. Rose, "Parameter selection for Newton-like methods applicable to nonlinear partial differential equations," *SIAM J. Numer. Anal.*, vol. 17, pp. 806-822, 1980.

[4] ——, "Global approximate Newton methods," *Numer. Math.*, vol. 37, pp. 279-295, 1981.

[5] R. E. Bank, J. W. Jerome, and D. J. Rose, "Analytical and numerical aspects of semiconductor device modeling," Bell Labs., Rep. 82-11274-2, 1982.

[6] G. Birkhoff, "The numerical solution of elliptic equations," in *Proc. SIAM*, Philadelphia, PA, 1971.

[7] F. J. Blatt, *Physics of Electronic Conduction in Solids.* New York: McGraw-Hill, 1968.

[8] E. M. Buturla, P. E. Cotrell, B. M. Grossman, K. A. Salsburg, M. B. Lawlor, and C. T. McMullen, "Three-dimensional finite element simulation of semiconductor devices," in *Proc. Int. Solid-State Circuits Conf.*, 1980, pp. 76-77.

[9] E. M. Buturla and P. E. Cotrell, "Simulation of semiconductor transport using coupled and decoupled solution techniques," *Solid-State Electron.*, vol. 23, pp. 331-334, 1980.

[10] E. M. Buturla, P. E. Cotrell, B. M. Grossman, and K. A. Salsburg, "Finite-element analysis of semiconductor devices: The FIELDAY program," *IBM J. Res. Dev.*, vol. 25, pp. 218-231, 1981.

[11] S. G. Chamberlain, A. Husain, "Three-dimensional simulation of VLSI MOSFET's," in *Proc. Int. Electron Devices Meeting*, 1981, pp. 592-595.

[12] P. E. Cotrell and E. M. Buturla, "Two-dimensional static and transient simulation of mobile carrier transport in a semiconductor," in *Proc. NASECODE I Conf.*, 1979, pp. 31–64.

[13] A. De Mari, "An accurate numerical one-dimensional solution of the p-n junction under arbitrary transient conditions," *Solid-State Electron.*, vol. 11, pp. 1021–2053, 1968.

[14] ——, "An accurate numerical steady-state one-dimensional solution of the p-n junction," *Solid-State Electron.*, vol. 11, pp. 33–58, 1968.

[15] P. Deuflhard, "A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with application to multiple shooting," *Numer. Math.*, vol. 22, pp. 289–315, 1974.

[16] P. Deuflhard and G. Heindl, "Affine invariant convergence theorems for Newton's method and extensions to related methods," *SIAM J. Numer. Anal.*, vol. 16, pp. 1–10, 1979.

[17] E. P. Doolan, J. J. H. Miller, and W. H. A. Schilders, *Uniform Numerical Methods for Problems with Initial and Boundary Layers.* Dublin: Boole Press, 1980.

[18] I. S. Duff, "A survey for sparse matrix research," *Proc. IEEE*, vol. 65, pp. 500–535, 1977.

[19] ——, "Practical comparison of codes for the solution of sparse linear systems," A.E.R.E. Harwell, Oxfordshire, 1979.

[20] T. Dupont, R. D. Kendall, and H. H. Rachford, "An approximate factorization procedure for solving self-adjoint elliptic difference equations," *SIAM J. Numer. Anal.*, vol. 5, pp. 559–573, 1968.

[21] W. L. Engl and H. Dirks, "Numerical device simulation guided by physical approaches," in *Proc. NASECODE I Conf.*, 1979, pp. 65–93.

[22] P. C. Fife, "Semilinear elliptic boundary value problems with small parameters," *Arch. Rat. Mech. Anal.*, vol. 29, pp. 1–17, 1973.

[23] L. Fox, "Finite-difference methods in elliptic boundary-value problems," in *The State of the Art in Numerical Analysis.* London: Academic Press, 1977, pp. 799–881.

[24] A. F. Franz, G. A. Franz, S. Selberherr, C. Ringhofer, and P. Markowich, "Finite boxes–A generalization of the Finite Difference method suitable for semiconductor device simulation," *IEEE Trans. Electron Devices*, vol. ED-30, pp. 1070–1082, 1983.

[25] J. Frey, "Physics problems in VLSI devices," in: *Introduction to the Numerical Analysis of Semiconductor Devices and Integrated Circuits.* Dublin: Boole Press, 1981, pp. 47–50.

[26] ——, "Transport physics for VLSI," in: *Introduction to the Numerical Analysis of Semiconductor Devices and Integrated Circuits.* Dublin: Boole Press, 1981, pp. 51–57.

[27] J. A. Greenfield, C. H. Price, and R. W. Dutton, "Analysis of nonplanar devices," in *NATO ASI on Process and Device Simulation for MOS-VLSI Circuits*, 1982.

[28] R. G. Grimes, D. R. Kincaid, and D. R. Young, "ITPACK 2A–A fortran implementation of adaptive accelerated iterative methods for solving large sparse linear systems," University of Texas, Austin, vol. CNA-164, 1980.

[29] H. K. Gummel, "A self-consistent iterative scheme for one-dimensional steady state transistor calculations," *IEEE Trans. Electron Devices*, vol. ED-11, pp. 455–465, 1964.

[30] K. Hess, "Ballistic electron transport in semiconductors," *IEEE Trans. Electron Devices*, vol. ED-28, pp. 937–940, 1981.

[31] W. Heywang and H. W. Pötzl, *Bandstruktur und Stromtransport.* Berlin: Springer, 1976.

[32] D. Kahng and M. M. Atalla, "Silicon–silicon dioxide field induced surface devices," *Solid-State Device Res. Conf.*, vol. IRE-AIEE, 1960.

[33] R. B. Kellog, "Analysis of a difference approximation for a singular perturbation problem in two dimensions," in *Proc. BAIL I Conf.*, Dublin: Boole Press, 1980, pp. 113–118.

[34] R. B. Kellog and H. Houde, "The finite element method for a singular perturbation problem using enriched subspaces," University of Maryland, Report BN-978, 1981.

[35] D. P. Kennedy and R. R. O'Brien, "Two-dimensional mathematical analysis of a planar type junction field-effect transistor," *IBM J. Res. Dev.*, vol. 13, pp. 662–674, 1969.

[36] J. Machek and S. Selberherr, "A novel finite-element approach to device modelling," *IEEE Trans. Electron Devices*, vol. ED-30, pp. 1083–1092, 1983.

[37] O. Manck and W. L. Engl, "Two-dimensional computer simulation for switching a bipolar transistor out of saturation," *IEEE Trans. Electron Devices*, vol. ED-24, pp. 339–347, 1975.

[38] A. H. Marshak and R. Shrivastava, "Law of the junction for degenerate material with position-dependent band gap and electron affinity," *Solid-State Electron.*, vol. 22, pp. 567–571, 1979.

[39] P. A. Markowich *et al.*, "An asymptotic analysis of single pn-junction devices," Mathematics Research Center, University of Wisconsin, Report 2527, 1983.

[40] P. A. Markowich, C. A. Ringhofer, S. Selberherr, and E. Langer, "A singularly perturbed boundary value problem modelling a semiconductor device," Mathematics Research Center, University of Wisconsin, Report 2388, 1982.

[41] P. A. Markowich *et al.*, "A singular perturbation approach for the analysis of the fundamental semiconductor equations," *IEEE Trans. Electron Devices*, vol. ED-30, pp. 1165–1180, 1983.

[42] P. A. Markowich, "Zur zweidimensionalen Analyse der Halbleitergrundgleichungen," Habilitation, Technical University of Vienna, 1983.

[43] M. S. Mock, "On equations describing steady-state carrier distributions in a semiconductor device," *Commun. Pure Appl. Math.*, vol. 25, pp. 781–792, 1972.

[44] ——, "A time-dependent numerical model of the insulated-gate field-effect transistor," *Solid-State Electron.*, vol. 24, pp. 959–966, 1981.

[45] C. Moglestue and S. J. Beard, "A particle model simulation of field effect transistors," in *Proc. NASECODE I Conf.*, 1979, pp. 232–236.

[46] C. Moglestue, "A Monte-Carlo particle model study of the influence of the doping profiles on the characteristics of field-effect transistors," in *Proc. NASECODE II Conf.*, 1981, pp. 244–249.

[47] A. Nussbaum, "Inconsistencies in the original form of the Fletcher boundary conditions," *Solid-State Electron.*, vol. 21, pp. 1178–1179, 1978.

[48] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables.* New York: Academic Press, 1970.

[49] C. H. Price, "Two-dimensional numerical simulation of semiconductor devices," Ph.D. dissertation, Stanford University, Stanford, CA, 1980.

[50] D. L. Scharfetter and H. K. Gummel, "Large-signal analysis of a silicon read diode oscillator," *IEEE Trans. Electron Devices*, vol. ED-16, pp. 64–77, 1969.

[51] A. Schütz, S. Selberherr, and H. W. Pötzl, "A two-dimensional model of the avalanche effect in MOS transistors," *Solid-State Electron.*, vol. 25, pp. 177–183, 1982.

[52] K. Seeger, *Semiconductor Physics.* Wien: Springer, 1973.

[53] S. Selberherr, A. Schütz, and H. W. Pötzl, "MINIMOS–A two-dimensional MOS transistor analyzer," *IEEE Trans. Electron Devices*, vol. ED-27, pp. 1540–1550, 1980.

[54] ——, "Two-dimensional MOS-transistor modeling," presented at NATO ASI on Process and Device Simulation for MOS–VLSI Circuits, 1982.

[55] G. D. Smith, *Numerical Solution of Partial Differential Equations: Finite Difference Methods.* Oxford: Clarendon Press, 1978.

[56] R. A. Smith, *Semiconductors.* Cambridge: Cambridge University Press, 1978.

[57] J. Stoer and R. Bulirsch, *Einführung in die Numerische Mathematik II.* Berlin: Springer, 1978.

[58] H. L. Stone, "Iterative solution of implicit approximations of multidimensional partial differential equations," *SIAM J. Numer. Anal.*, vol. 5, pp. 530–558, 1968.

[59] T. Toyabe, K. Yamaguchi, S. Asai, and M. Mock, "A numerical model of avalanche breakdown in MOSFET's," *IEEE Trans. Electron Devices*, vol. ED-25, pp. 825–832, 1978.

[60] R. J. Van Overstraeten, H. J. De Man, and R. P. Mertens, "Transport equations in heavy doped silicon," *IEEE Trans. Electron Devices*, vol. ED-20, pp. 290–298, 1973.

[61] W. V. Van Roosbroeck, "Theory of flow of electrons and holes in germanium and other semiconductors," *Bell Syst. Tech. J.*, vol. 29, pp. 560–607, 1950.

[62] K. M. Van Vliet, "On Fletcher's boundary conditions," *Solid-State Electron.*, vol. 22, pp. 443–444, 1979.

[63] ——, "The Shockley-like equations for the carrier densities and the current flows in materials with a nonuniform composition," *Solid-State Electron.*, vol. 23, pp. 49–53, 1980.

[64] R. S. Varga, *Matrix Iterative Analysis.* Englewood Cliffs, NJ: Prentice-Hall, 1962.

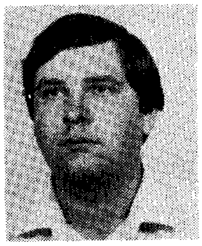
[65] A. B. Vasilev'a and V. F. Butuzov, "Singularly perturbed equations in the critical case," Mathematics Research Center, University of Wisconsin, translated Report 2039, 1978.

[66] E. L. Wachspress, *Iterative Solution of Elliptic Systems.* Englewood Cliffs, NJ: Prentice-Hall, 1966.

[67] A. Yoshii, S. Horiguchi, and T. Sudo, "A numerical analysis for very small semiconductor devices," in *Proc. Int. Solid-State Circuits Conf.*, 1980, pp. 80-81.

[68] A. Yoshii, H. Kitazawa, M. Tomizawa, S. Horiguchi, and T. Sudo, "A three-dimensional analysis of semiconductor devices," *IEEE Trans. Electron Devices*, vol. ED-29, pp. 184-189, 1982.

*



Siegfried Selberherr (M'79) was born in Klosterneuburg, Austria, on August 3, 1955. He received the degree of "Diplomingenieur" in control theory and industrial electronics from the Technical University of Vienna, Vienna, Austria in 1978. He completed his dissertation on "Two dimensional MOS-transistor modeling" in 1981.

Upon graduation he joined the "Institut für Allgemeine Elektrotechnik and Elektronik"—previously called the "Institut für Physikalische Elektronik"—at the Technical University of Vienna as an Assistant Professor. His current topics of interest are modeling and simulation of devices and circuits for application in electronic systems.

Dr. Selberherr is a member of the Association for Computing Machinery and the Society of Industrial and Applied Mathematics.

*

Christian A. Ringhofer was born in Winnipeg, Canada, on February 9, 1957. He attended the University of Vienna, Vienna, Austria, from 1975 to 1976 studying law. From 1976 to 1981 he attended the Technical University of Vienna, where he received the Diplom-Ingenieur degree in 1980 and the Dr. techn. in 1981.

He worked as a Research Associate at the Institute for Applied Mathematics, Technical University of Vienna from 1980 to 1982, and he has since been a Research Associate at the Mathematics Research Center, University of Wisconsin, Wisconsin, Madison. His fields of interest include numerical analysis, partial differential equations, and applied mathematics.

Three-Dimensional Monte Carlo Simulations—Part I: Implanted Profiles for Dopants in Submicron Device

A. M. MAZZONE AND G. ROCCA

Abstract—**Monte Carlo methods are used to simulate implants. The results fall into two different groups. On one side, size-dependent effects due to the presence of the mask are analyzed and discussed. On the other side, physical mechanisms dependent on dose and energy, like channeling and transition crystal-amorphous, are briefly reviewed.**

I. Introduction

There is a continuous trend towards scaled-down devices, and MOS with gatelength of a few thousands angstroms are actively being studied almost everywhere.

If one considers that 1000 Å represents approximately two hundred atomic planes, it is plausible to say that process modeling must incorporate, today or in the near future, the methods traditionally used for lattice studies, like Monte Carlo and molecular dynamics.

These methods, defined "computer experiments" by G. H. Vinejard, who lead the modern school of lattice simulation, circumvent the analytical difficulties connected with transport and percolation problems treating the given case without any *a priori* assumption. Though plagued by a limited knowledge of atomic parameters, the resulting picture is rarely far from reality.

In the case of ion implantation, the current means of analytical evaluation are based on statistical approaches which regard the target as homogeneous and amorphous. These methods have reached a high degree of sophistication and the assumption of a homogeneous target has been actually removed. Smith and Gibbons [1] used a semi-analytical method to solve in one dimension the linear Boltzmann equation for a multilayered target and Christel and Gibbons [2] extended the method to the ion-beam induced interface mixing. Apart from these works, statistical approaches generally lead to an accurate evaluation of the moments of the ion range and of the deposited energy distribution. However, an arbitrary choice for the corresponding distribution function remains.

Manuscript received August 8, 1983.

The authors are with CNR Istituto LAMEL, Via Castagnoli, 1-40126, Bologna, Italy.

0028-0070/84/0100-0064$01.00 © 1984 IEEE