Nonlinear Multigrid applied to a 1D Steady Transistor Problem

P.M. de Zeeuw

CWI, Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

1. Introduction

In this paper we restrict ourselves on purpose to one space dimension as a preparatory study for the case or more space dimensions. We study a particular example problem which has been put forward by Dr. W.H.A. Schilders, Philips, The Netherlands. This problem models a transistor and turns out to be a lot harder to solve than the forward or reversed biased diode problem. We apply nonlinear multigrid and encounter a serious difficulty due to the nonlinearity of the problem. The difficulty is analysed and this provides insight equally well for the case of more space dimensions. Some modifications are proposed which significantly increase the robustness of the nonlinear multigrid method and which look promising also for the more-dimensional case.

2. The problem

The behaviour of a steady semiconductor device can be described by the following set of equations (cf. e.g. [5]):

$$-\nabla J_\psi = n_i q(exp(\alpha(\phi_p-\psi)) - exp(\alpha(\psi-\phi_n))) + qD, \tag{2.1a}$$

$$-\nabla J_n = +qR, \tag{2.1b}$$

$$-\nabla J_p = -qR, \tag{2.1c}$$

where J_ψ is defined by

$$J_\psi = \varepsilon\nabla\psi, \tag{2.2a}$$

and

$$J_n = \bar{\mu}_n exp(\alpha(\psi-\phi_n))\nabla(\alpha\phi_n), \tag{2.2b}$$

$$J_p = \bar{\mu}_p exp(\alpha(\phi_p-\psi))\nabla(\alpha\phi_p), \tag{2.2c}$$

with

$$\bar{\mu}_n = \frac{n_i q\mu_n}{\alpha},\ \bar{\mu}_p = \frac{n_i q\mu_p}{\alpha}.$$

Substitution of (2.2) into (2.1) results in a system of three nonlinear partial differentaial equations for the variables

$$\psi,\ \phi_n \text{ and } \phi_p.$$

In (2.1) ψ represents the electrostatic potential, ϕ_n and ϕ_p are the hole and electron quasi-Fermi potentials. On the one hand, by choosing these variables, the problem is strongly nonlinear, on the other hand the values assumed by (ψ, ϕ_n, ϕ_p) are in a moderate range. The quasi-Fermi potentials satisfy the well-known relations:

$$n = n_i exp(\alpha(\psi-\phi_n)), \tag{2.3a}$$

$$p = n_i exp(\alpha(\phi_p - \psi)), \qquad (2.3b)$$

where p and n describe the concentration of holes and electrons respectively. Equations (2.1b) and (2.1c) are called the continuity equations; J_n is the electron current density, J_p is the hole current density, R is the recombination-generation rate, a function of n and p. The doping profile D is a function of the space variable x. The quantities ε, q, α, μ_n, μ_p represent the permittivity, the elementary charge, the inverse of the thermal voltage and the electron and hole mobility respectively. In this paper we consider the case of only one space dimension and assume ε, α, μ_n and μ_p to be constant.

2.1 A particular 1D model problem.

We will focus our attention to a particular (hard) 1D model problem which has been supplied by Dr. W.H.A. Schilders (cf.[6,7]). Here the problem constants are

$$\varepsilon = 1.035918_{10}^{-12}, q = 1.6021_{10}^{-19}, \mu_n = \mu_p = 500, n_i = 1.22_{10}^{10},$$

$$\alpha = q/KT, k = 1.38054_{10}^{-23}, T = 300. \qquad (2.4)$$

The function R is given by

$$R = \frac{pn - n_i^2}{\tau(p + n + 2n_i)}, \quad \tau = 10^{-6}.$$

In figure 2 a graph of the doping function $D(x)$ is shown after the transformation

$$D \rightarrow sign(D)\ ^{10}log(1 + |D|).$$

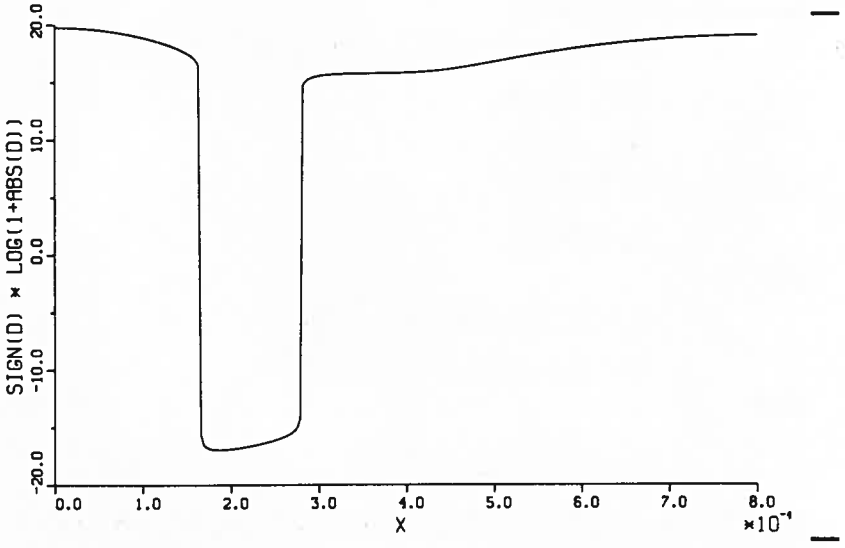


FIGURE 1. The doping profile.

The problem is defined on the domain $\Omega = [0, 8_{10}^{-4}]$. We have three contacts to our semiconductor device (the 1D-model of a transistor):

the emitter E at $x = 0$,

the basis B at $x = 1.9_{10}^{-4}$,

the collector C at $x = 8_{10}^{-4}$.

Boundary conditions at the emitter are:

$p - n + D = 0$ (i.e. vanishing space charge),

$\phi_n = V_E$,

$J_p = 0$.

Boundary conditions at the basis:

$\phi_p = V_B = 0$.

Boundary conditions at the collector:

$p - n + D = 0$,

$\phi_n = V_C$,

$\phi_p = V_C$.

For fifteen different cases, each characterized by a pair of voltages (V_E, V_C), the solution is required.

case	V_E	V_C
0	0.	0.
1	0.	0.2
2	0.	0.4
3	0.	0.6
4	0.	0.8
5	0.	1.
6	-0.2	1.
7	-0.4	1.
8	-0.6	1.
9	-0.7	1.
10	-0.8	1.
11	-0.85	1.
12	-0.9	1.
13	-0.95	1.
14	-1.	1.

TABLE 2.1. Subsequent voltages at the emitter and collector for which a solution is required.

3. The discretization

The interval $\Omega = (0, 8_{10}^{-4})$ is split up into disjoint *boxes* $B_j = (x_{j-1}, x_j)$, $j = 1\ (1)\ N$.

A point x_j is called a *wall*, a point $x_{j-1/2} = (x_{j-1} + x_j)/2$ is called a *center*. Another set of subintervals $\{D_j\}$ is defined by

$D_0 = (x_0, x_{1/2})$,

$D_j = (x_{j-1/2}, x_{j+1/2})$,

$D_N = (x_{N-1/2}, x_N)$.

This set is called the set of *dual boxes*. The following conditions should be satisfied:

(i) $x_{j-1} < x_j$,

(ii) $x_0 = E$, $x_N = C$,

(iii) $\exists j^*$ with $0 < j^* < N$ such that $x_{j^*} = B$, i.e. the basis B is at the partition-wall between two boxes.

By using a finite volume technique we arrive at a cell-centered version of the well-known Scharfetter-Gummel scheme (cf. e.g.[4,7]). Firstly we apply the Gauss divergence theorem in one dimension to the equations (2.1) on the domains B_j. Secondly we use the assumption that J_ψ, J_n, and J_p are piecewise constant on the dual set $\{D_j\}$ (cf. e.g.[1]); therefore we introduce the notation

$$J_{\psi\,j},\ J_{n\,j},\ \text{and}\ J_{p\,j}.$$

At the basis B special measures are taken, for full details cf.[7]. Further we introduce the variables

$$(\psi_j, \phi_{n\,j}, \phi_{p\,j})^T,\ j = 1\ (1)\ N,$$

which are associated with the centers of the boxes B_j. In this way we arrive at the following set of discrete equations:

$$-J_{\psi\,j} + J_{\psi\,j-1} - S_j = F_j,\ j = 1\ (1)\ N, \tag{3.1a}$$

$$-J_{n\,j} + J_{n\,j-1} - R_j = 0,\ j = 1\ (1)\ N, \tag{3.1b}$$

$$-J_{p\,j} + J_{p\,j-1} + R_j = 0,\ j = 1\ (1)\ N,\ j \neq j^*,\ j \neq j^*+1. \tag{3.1c}$$

The piecewise constant functions are derived by (cf.[4])

$$J_{\psi\,j} = \varepsilon \frac{\psi_{j+1}-\psi_j}{x_{j+1/2}-x_{j-1/2}}, \tag{3.2a}$$

$$J_{n\,j} = \bar{\mu}_n \frac{exp(-\alpha\phi_{n\,j+1})-exp(-\alpha\phi_{n\,j})}{exp(-\alpha\psi_{j+1})-exp(-\alpha\psi_j)} \cdot \frac{\alpha\psi_{j+1}-\alpha\psi_j}{x_{j+1/2}-x_{j-1/2}}, \tag{3.2b}$$

$$J_{p\,j} = \bar{\mu}_p \frac{exp(\alpha\phi_{p\,j+1})-exp(\alpha\phi_{p\,j})}{exp(\alpha\psi_{j+1})-exp(\alpha\psi_j)} \cdot \frac{\alpha\psi_{j+1}-\alpha\psi_j}{x_{j+1/2}-x_{j-1/2}}. \tag{3.2c}$$

The other functions are defined by:

$$F_j \approx q \int_{B_j} D\,d\Omega, \tag{3.3a}$$

$$S_j \approx n_i q \int_{B_j} (exp(\alpha(\phi_p-\psi))-exp(\alpha(\psi-\phi_n)))\,d\Omega, \tag{3.3b}$$

$$R_j \approx q \int_{B_j} R\,d\Omega, \tag{3.3c}$$

where $\approx$ denotes approximation by midpoint quadrature. For full details on the discretization of the boundary conditions cf.[7]. In this way we have now discretized the equations (2.1) into a set of 3N nonlinear equations (3.1). We can write (3.1) in symbolic form as

$$M_h(q_h) = f_h, \tag{3.4}$$

where M_h denotes the nonlinear difference operator corresponding with the lefthandside of (3.1) and where f_h corresponds with the righthandside of (3.1) and q_h with the variables $(\psi_j, \phi_{n\,j}, \phi_{p\,j})^T$.

3.1. Properties of the Jacobian

In this subsection we study how the Jacobian of M_h depends on the discrete solution. We confine ourselves to the dependency on ϕ_p, results for ϕ_n can be derived analogously. First we assume the recombination term to be zero. We freeze the solution components ψ and ϕ_n and consider the ϕ_p-stencil, at box B_j, defined by the triplet

$$[st_p(j,-1), st_p(j,0), st_p(j,+1)] \tag{3.5a}$$

with

$$st_p(j,k) = \frac{\partial(-J_{p\,j}+J_{p\,j-1})}{\partial \phi_{p\,j+k}},\ k = -1,\ 0,\ 1. \tag{3.5b}$$

Because of

$$st_p(j,-1)<0,\ st_p(j,0)>0,\ st_p(j,+1)<0 \tag{3.6}$$

the stencil corresponds with an *l*-matrix.
In ref.[7] the following observation is made upon this stencil:

$$st_p(j,0) = -(st_p(j,-1) + st_p(j,+1)) + \alpha(-J_{p\,j}+J_{p\,j-1}). \tag{3.7}$$

It follows from (3.7) that at the discrete solution, i.e. when

$$-J_{p\,j}+J_{p\,j-1} = 0$$

is satisfied, the following equality holds:

$$st_p(j,0) = -(st_p(j,-1) + st_p(j,+1)),$$

so then the *l*-matrix possesses also weak diagonal dominance. However, in the middle of some iterative process to determine the solution, we may well have negative residuals so that the equality (3.7) implies the loss of diagonal dominance. Therefore ill-conditioning and numerical difficulties can be expected. (If the recombination term is not neglected then $st_p(j,0)$ is enlarged with some positive value.) This is one explanation for the difficulties encountered when applying the Newton method.

4. The Newton Method and Expedients

Applying the Newton method directly to (3.4) involves large storage requirements and the solution of large linear systems (it would extremely so in two or more space dimensions). Therefore it should be applied only for relatively coarse grids. In fact we will use it as a *coarsest grid solver* for multigrid methods which will be treated in the next section. We use some additional tools to enhance the global convergence of the Newton method:

1) correction transformation cf.[5],
2) continuation (with boundary voltages as parameter) cf. e.g. [6],
3) smoothing of the Newton-iterates by means of Collective Symmetric Gauss Seidel relaxation (CSGS) cf.[4].

5. The Multigrid Method

Advanced ways of solving a set of nonlinear equations are the Full Approximation Scheme (FAS), cf.[2] and the Nonlinear Multigrid Method (NMGM), cf.[3]. Both multigrid methods are very similar although the NMGM is more general. Recently, in the field of semiconductor equations research on multigrid methods has been initiated. If well applied, a multigrid method can be optimal in the sense that the rate of convergence is independent of the meshsize. An important advantage of the FAS/NMGM-method is that no large linear systems need to be stored and solved. The subsequent stages of an usual FAS-method, applied to (3.4), are:

1) apply p nonlinear relaxation sweeps; thus we get an approximation q_h of the solution which has a smooth residual $d_h \equiv f_h - M_h(q_h)$,

2a) transfer d_h from Ω_h to a coarser grid Ω_H by means of the restriction operator R_H, and choose a representation q_H^{old} on the coarse grid of q_h on the fine grid,

2b) solve (approximately) on Ω_H the equation

$$M_H(q_H^{new}) = M_H(q_H^{old}) + \bar{R}_H d_h, \tag{5.1}$$

2c) prolongate the correction, computed on Ω_H, onto Ω_h and add to q_h:

$$q_h^{new} = q_h + (P_h q_H^{new} - P_h q_H^{old}), \tag{5.2}$$

3) apply q nonlinear relaxation sweeps.

Stage 2 is called the coarse grid correction (CGC). Stage 2b may be obtained by applying a number of σ FAS-cycles on the coarser grid. In this way a recursive procedure is obtained in which a sequence of increasingly coarser grids is used. In this paper we use $p = q = \sigma = 1$ throughout. As relaxation we use CSGS. Let the coarsest grid Ω_H, a discretization of Ω, be given by the set of boxes $\{B_{H\,j}\}$. From Ω_H we construct the next finer grid Ω_h by division of each box $B_{H\,j}$ into two disjoint boxes $B_{h\,2j-1}$ and $B_{h\,2j}$. By repetition we obtain thus a sequence of increasingly finer grids. For the numerical experiments in this paper we assume in addition that $B_{h\,2j-1}$ and $B_{h\,2j}$ have equal size. The restriction operator $\bar{R}_H$ for right hand side functions is defined by adding two adjacent values:

$$(\bar{R}_H f_h)_j = f_{h\,2j-1} + f_{h\,2j}. \tag{5.3}$$

As a prolongation operator P_h which transfers a solution from a coarse grid to the next finer one we use a prolongation introduced by Hemker cf.[4] which is based on the assumption of smoothness of fluxes and which satisfies the Galerkin condition. Various possibilities exist for choosing the approximation q_H^{old}, e.g. the application of fullweighting to q_h. Another possibility is to take q_H^{old} equal to q_H^{new} obtained from the previous CGC. In practice, the solution efficiency of nonlinear problems was never shown to be much influenced by either choice of q_H^{old}, in our case however it is (see Section 6).

6. Adaptation of the coarse grid correction

Hemker successfully applied boxcentered multigrid FAS iteration to the forward and the reversed biased diode problem (cf.[4]). However, application of the same algorithm to the transistorproblem in this paper gives rise to a complication in the CGC due to drastically varying problem coefficients. This complication and possible remedies are the topics of this section.

6.1 Improper solution transfer

The first attempt of applying multigrid to our specific problem was done by employing FAS, embedded within the Full Multigrid Method (FMG), with only two grids. The coarse grid problem within the CGC of FAS, was to be solved up to machine-accuracy by means of the Newton method with expedients (see Section 4). For several cases of the testproblem it turned out that the twogrid-algorithm gets stuck precisely at this stage. This is remarkable because the Newton method (with expedients) on his own is successful even for very coarse grids. Apparently the right hand side of the equation (5.1) in stage 2b) of the CGC, may be outside a proper range of M_H. Whenever such a difficulty

occurred one or more of the three solution components depicted a steep gradient and indeed, in a way, this is causing the trouble.

For an explanation, again consider the ϕ_p-stencil (see section 3.1). Let the operator Δ_j denote the variation of a gridfuntion over the centers of two adjacent boxes B_j and B_{j-1}. In [7] it is shown that if both $|\Delta_{j-1}\psi|$ and $|\Delta_j\psi|$ are sufficiently small then

$$st_p(j,0) \approx \alpha\bar{\mu}_p exp(\alpha(\phi_{p\,j}-\psi_j))\cdot(\frac{1}{\Delta_{j-1}x}+\frac{1}{\Delta_j x}), \tag{6.1}$$

and if both $|\Delta_{j-1}\psi|$ and $|\Delta_j\psi|$ are sufficiently large then

$$st_p(j,0) \approx \alpha\bar{\mu}_p exp(\alpha(\phi_{p\,j}-\psi_j))\cdot \begin{cases} \alpha\frac{\Delta_{j-1}\psi}{\Delta_{j-1}x} & \text{if } \Delta_{j-1}\psi\geq 0, \\ -\alpha\frac{\Delta_j\psi}{\Delta_j x} & \text{if } \Delta_j\psi\leq 0. \end{cases}$$

These approximations show that the stencil is extremely sensitive to the difference $(\phi_{p\,j}-\psi_j)$. Hence the ϕ_p-stencil on the coarse grid is extremely sensitive to how $\phi_{p\,j}$ and ψ_j on the coarse grid are determined from their counterparts on the fine grid. Because of a steep gradient it may well occur that the problem coefficients, i.e. the entries of the Jacobian of M_h, show a quite different order of magnitude on two adjacent boxes. Therefore, on the coarse grid, the problem coefficients do heavily depend on the particular coarse grid representation q_H of q_h. When this particular coarse grid representation q_H generates small problem coefficients, the righthand side of equation (5.1) may easily become out of the range of M_H (note that $R_H d_h$ does not depend on the particular choice of q_H). We conclude that at locations where such a phenomenon occurs we cannot expect to be able to construct a coarse grid operator M_H which is a fair representation of the fine grid operator M_h. For a more detailed discussion cf.[7].

6.2 Possible remedies

A radical remedy to meet the above sketched difficulty is to prevent the variation $\Delta_j(\alpha(\phi_p-\psi))$ over two adjacent boxes from getting large, i.e. to introduce local refinement of the mesh by equidistributing the variation. However, we want to be able to apply coarse grids in our multigridmethod. Therefore we resort to another remedy.

Let L and R be the centers of two adjacent boxes on the fine grid Ω_h which together constitute a box with center M on the coarse grid Ω_H (see figure 2).

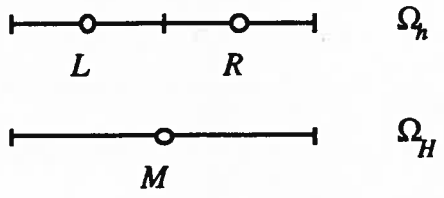


FIGURE 2. Nested boxes.

The centers of the ϕ_p-stencils at L, R and M are the coefficients a_L^h, a_R^h and a_M^H respectively. Let $d_h(L)$ and $d_h(R)$ be the residuals at L, R (e.g. for the third equation only). At M the difference between q_H^{new} and q_H^{old} has the order of magnitude $d_h(M)/a_M^H$ with

$$d_H(M) = d_h(L) + d_h(R),$$

see equation (5.1). Because of a steep gradient in the solution it may occur that

$$d_M^H << \max\{a_L^h, a_R^h\},$$

which implies that the coarse grid correction prolongated to the fine grid becomes far too large and the solution q_h gets spoiled. A way to prevent this situation is to multiply the restricted residual $d_H(M)$ with

$$\theta_M^H \equiv \frac{d_M^H}{\max\{a_L^h, a_R^h\}}, \quad 0<\theta_M^H \leq 1. \tag{6.2}$$

For a smooth part of the solution this fraction will be near to 1, for a rapidly varying part of the solution it will be near to 0 so that the old solution q_h will be preserved (see (5.2)). Note that θ_M^H is a different number for each box on the coarse grid. Note also that for increasingly fine grids all numbers θ_M^H become 1. With this local suppression of the restricted residual our FAS-algorithm has been modified. It has been applied at *all coarser grids* of the nonlinear multigridalgorithm for the continuity-equations. A technical but noteworthy detail is that in stead of θ_M^H, the number

$$\theta_M^{H\,\prime} \equiv \min\{2\theta_M^H, 1\}, \tag{6.3}$$

has been used as suppression-factor in the numerical experiments.
For more details cf.[7]. Numerical results are shown in the next section.

7. Numerical results

In this section we investigate the performance of our nonlinear multigridalgorithm. We focus our attention on the effects of local suppression of the restricted residual and the particular choice of the coarse grid solution q_H^{old}. In table 7.1 the performance of the nonlinear multigrid algorithm is measured by the average number of sweeps necessary to obtain a reduction factor 10^{-1} of the maximumnorm of the residual. In the heading of the table the following abbreviation are used:

case	: this refers to table 2.1,
no θ	: no suppression-operator for the restricted residual has been applied,
θ yes	: the suppression-operator for the restricted residual has been applied,
q_H^{old} etc.	: defines how q_H^{old} in (5.1) has been chosen.

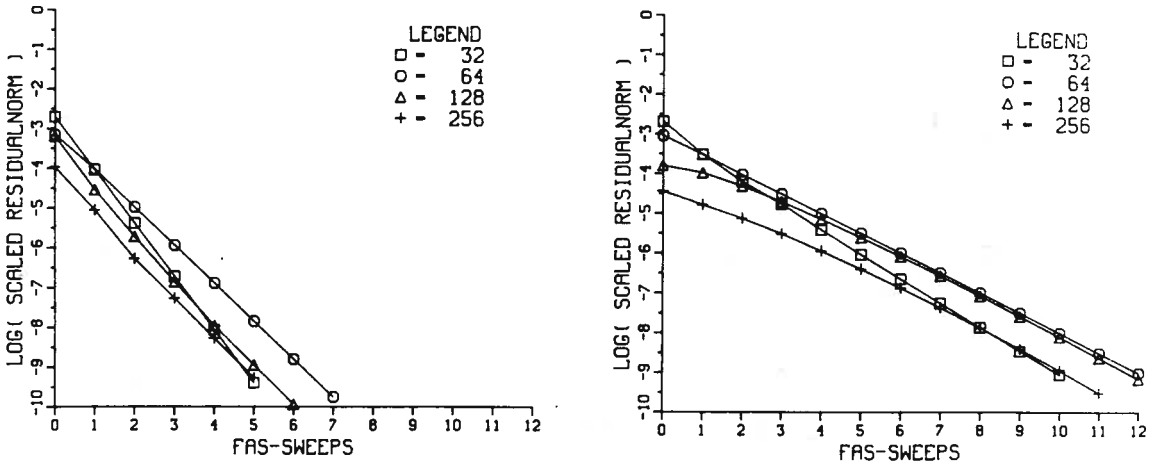
In the event of divergence we write *div.

	$q_H^{old} : q_H^{new}$ of previous CGC		q_H^{old} : by fullweighting on q_h	
case	no θ	θ yes	no θ	θ yes
0	0.66	0.66	0.80	0.80
1	0.88	0.73	1.07	1.07
2	*div	0.85	*div	1.15
3	*div	1.03	*div	*div
4	*div	1.06	*div	*div
5	*div	0.89	*div	*div
6	*div	0.89	*div	*div
7	*div	0.89	*div	1.38
8	0.89	0.89	1.40	1.37
9	0.87	0.90	1.54	1.37
10	0.88	0.82	2.59	2.52
11	1.25	1.25	1.22	1.22
12	2.01	2.01	1.75	1.74
13	2.19	2.19	4.66	4.66
14	1.76	1.77	2.44	2.43

TABLE 7.1. Performance of FAS, average number of sweeps necessary to obtain a reduction factor 10^{-1} of the maximumnorm of the residual; 3 grids: N =16, 32, 64 respectively.

Table 7.1 shows that the use of the θ-operator is not sufficient on his own to guarantee convergence, apparently the full weighting approximation of the fine grid solution on the coarse grid may be a poor one (case 3 - 6). The use of the θ-operator combined with a proper choice of the coarse grid solution gives convergence for all cases. The use of the θ-operator does not slow down convergence in the cases where it is not needed.

For two typical cases, case = 4 and case = 12, we investigate the grid-dependence of the multigrid convergence in detail. In figure 3 we show the 10-logarithm of the scaled residual norms after subsequent FAS-sweeps, starting from the result by fullmultigrid with one FAS-sweep at each level. The coarsest grid always contains 16 boxes, for the finest grid we take 32, 64, 128 and 256 boxes respectively. Thus we apply multigrid with 2, 3, 4 and 5 grids respectively. The local suppression of the restricted residuals (θ-operator) has been applied at all coarser grids.



a. case = 4.

b. case=12.

FIGURE 3. Multigrid convergence histories; the coarsest grid numbers 16 boxes.

We observe that the multigrid convergence becomes grid-independent for increasingly finer grids.

8. Conclusions

At first, when applying the nonlinear multigrid to our 1D-transistorproblem we encounter several cases with severe divergence. The difficulty is caused by discrepancies among the discretizations of the nonlinear differential operator on the subsequent grids, due to the rapidly varying problem coeffficients. This difficulty is met by adaptation of the coarse grid correction. A proper choice of the coarse grid solution is of importance too, the full weighting approximation is not satisfactory. With the improvements as proposed we obtain fast multigrid convergence with a convergence rate that is independent of the meshsize.

References

[1]. R.E.Bank, D.J.Rose, W.Fichtner (1983). *Numerical methods for semiconductor device simulation*. SIAM J. Sci. Stat. Comput., Vol.4, No 3, pp.416-435.

[2]. A.Brandt (1981). *Guide to multigrid development*. Lecture notes in Mathematics 960, W. Hackbusch and U.Trottenberg, eds., Springer-Verlag, Berlin, pp.220-312.

[3]. W.Hackbusch. (1985). *Multigrid methods and applications*. Springer Series in Computational Mathematics 4, Springer Verlag, Berlin.

[4]. P.W. Hemker. (1988). *A multigrid approach for one-dimensional semiconductor device simulation*. GMD-Studien Nr. 144 on Device Simulation. W.Joppich and U.Trottenberg, eds., GMD, Sankt Augustin.

[5]. S.J. Polak, C. den Heijer, W.H.A. Schilders, P.A. Markowich. (1987). *Semiconductor device modelling from the numerical point of view*. Int. J. Numer. Math. Engineering 24, pp. 763-838.

[6]. B.P. Sommeijer, W.H. Hundsdorfer, C.T.H. Everaars, P.J. van der Houwen, J.G.Verwer. (1987). *A numerical study of a 1D stationary semiconductor model*. Note NM-8702, Dept. of Numerical Mathematics, Centre for Mathematics and Computer Science, P.O. BOX 4079, 1009 AB Amsterdam, The Netherlands.

[7]. P.M. de Zeeuw. (1989). *Nonlinear multigrid applied to a 1D stationary semiconductor model*. Report NM-R8905, Dept. of Numerical Mathematics, Centre for Mathematics and Computer Science, P.O. BOX 4079, 1009 AB Amsterdam, The Netherlands.

GMD-Studien Nr. 177

Wolfgang Joppich (Hrsg.)

Device Simulation

2. Workshop in der GMD
15.–16. November 1989

März 1990

GESELLSCHAFT FÜR MATHEMATIK
UND DATENVERARBEITUNG MBH

P.M. de Zeeuw