Fast Iterative Solution of Carrier Continuity Equations in 3D MOS/MESFET Simulations

O. Heinreichsberger
S.Selberherr
M.Stiftinger

Institut für Mikroelektronik
Technische Universität Wien
Gusshausstrasse 27–29, A –1040 Wien, Austria

February, 1990

Abstract

We summarize our experience in solving the three-dimensional carrier continuity equations in our MOS/MES–FET simulator MINIMOS 5. First we give a brief overview of the algebraic properties of the coefficient matrices. We show that the matrices are symmetrizable and can be solved by the symmetrized preconditioned CG algorithm. Since the symmetrization matrices are computationally infeasible due to their enormous dynamical range, we turn our focus to various iterative accelerated methods for nonsymmetric matrices. Of these methods the BICGS algorithm together with (unmodified) ILU preconditioning exhibits an optimum of reliability, convergence speed and memory consumption. A controllable level of fill-in during factorization can handle the badly conditioned systems which we frequently find in our simulations.

1 The Basic Partial Differential Equations

In our MOS/MES–FET simulator MINIMOS 5 the static semiconductor equations are solved self-consistently on a three-dimensional rectangular domain using finite difference discretization. The static semiconductor equations [18] for the variables (ψ, n, p) consist of the Poisson equation

$$div\,(\epsilon \cdot grad\psi) = -\rho$$

with the space charge $\rho = q \cdot (p - n + N_D^+ - N_A^-)$, and of the carrier continuity equations

$$\begin{aligned} div\vec{J_n} &= q \cdot R \\ div\vec{J_p} &= -q \cdot R \end{aligned}$$

where the carrier transport is modelled by an extended drift–diffusion approach

$$\begin{aligned} \vec{J_n} &= q \cdot \mu_n \cdot n \cdot \left(grad\psi + \tfrac{1}{n} \cdot grad\left(n \cdot \tfrac{k \cdot T_n}{q}\right)\right) \\ \vec{J_p} &= q \cdot \mu_p \cdot p \cdot \left(grad\psi - \tfrac{1}{p} \cdot grad\left(p \cdot \tfrac{k \cdot T_p}{q}\right)\right) \end{aligned}$$

Let $F_{n,p}$ denote the effective driving forces for electrons and holes that depend on the local carrier temperatures $T_{n,p}$ [3][19]

$$\begin{aligned} \vec{F_n} &= \left|grad\psi - \tfrac{1}{n} \cdot grad\left(\tfrac{k \cdot T_n}{q} \cdot n\right)\right| \\ \vec{F_p} &= \left|grad\psi + \tfrac{1}{p} \cdot grad\left(\tfrac{k \cdot T_p}{q} \cdot p\right)\right| \end{aligned}$$

Carrier heating is modelled by a carrier temperature depending mobility $\mu_{n,p}^{LISF}$

$$\mu_{n,p}^{LISF} = \frac{2\mu_{n,p}^{LIS}}{1 + \left(1 + \left(\frac{2\mu_{n,p}^{LIS} F_{n,p}}{v_{n,p}^{sat}}\right)^{\alpha_{n,p}}\right)^{\frac{1}{\alpha_{n,p}}}}$$

with $\alpha_n = 2, \alpha_p = 1$, where $\mu_{n,p}^{LIS}$ denotes the zero field mobility due to lattice–, impurity- and surface scattering [19][20]. The carrier temperatures $T_{n,p}$ are formulated as

$$T_{n,p} = T_0 + \frac{2}{3} \cdot \frac{q}{k} \cdot \tau_{n,p} \cdot \left(v_{n,p}^{sat}\right)^2 \cdot \left(\frac{1}{\mu_{n,p}^{LISF}} - \frac{1}{\mu_{n,p}^{LIS}}\right)$$

The expression R on the right hand side of the transport equations represents the sum of the impact ionization rate, the Shockley–Read–Hall and Auger recombination rates

$$R = R^{II} + R^{SRH} + R^{AU}$$

Impact ionization is modelled by [18]

$$R^{II} = -\alpha_n \cdot \frac{\left|\vec{J_n}\right|}{q} - \alpha_p \cdot \frac{\left|\vec{J_p}\right|}{q}$$

in which the $\alpha_{n,p}$ depend exponentially on the local absolute electric field

$$\alpha_{n,p} = \alpha_{n,p} \cdot exp\left(-\frac{\beta_{n,p}}{|E|}\right)$$

The SRH recombination rate is expressed by

$$R^{SRH} = \frac{\left(n \cdot p - n_i^2\right)}{\tau_p\left(n + n_1\right) + \tau_n\left(p + p_1\right)}$$

and the Auger recombination rate is given by

$$R^{AU} = \left(C_n \cdot n + C_p \cdot p\right) \cdot \left(n \cdot p - n_i^2\right)$$

2 Discretization and Iterative Solution of the Nonlinear System of Equations

In MINIMOS the set of equations described in the previous section is discretized on a threedimensional rectangular domain together with appropriate boundary conditions. For the idealized ohmic contacts we have Dirichlet boundary conditions. For the artificial interfaces in the deep bulk homogenous Neumann boundary conditions are applied, whereas for the case of non-vanishing interface charge and interface recombination velocity we have non-homogenous Neumann boundary conditions. We note that the boundary conditions may be given implicitly. For example in the case of the Schottky contact in a MESFET simulation with MINIMOS, the boundary condition is given by the current densities at the Schottky contact [11]

$$\begin{aligned} J_n &= -q \cdot v_n \cdot (n - n_0) \\ J_p &= q \cdot v_p \cdot (p - p_0) \\ n_0 &= n_i \cdot exp\left(-\frac{\psi_s}{U_t}\right) \\ p_0 &= n_i \cdot exp\left(\frac{\psi_s}{U_t}\right) \end{aligned}$$

where ψ_s is the (fixed) surface potential on the Schottky contact. The nonlinear system of equations is solved by decoupling the three partial differential equations (Gummel's algorithm [1]). A tensor product grid, strongly nonuniform due to the rapid variation of the carrier concentrations, is used. MINIMOS 5 is capable of handling non-planar interfaces. This is achieved via the well known box integration concept.

The nonlinear Poisson equation is linearized by a one-step Newton iteration. Using finite difference discretization, the resulting system of linear equations is symmetric, positive definite and has property A.

The carrier continuity equations are nonlinear as well, due to the dependence of $\mu_{n,p}$ on the driving forces and of the right hand side R on ψ, n and p.

The dependence of the mobilities $\mu_{n,p}$ on n and p, respectively, is neglected. The influence of the derivatives of the impact ionization rate with respect to the carrier concentrations is neglected, too. This seems justified, since this quantity is not updated at every nonlinear iteration but in a generation *subcycle*. For the recombination rates R^{SRH} and R^{AU}, however, the derivatives with respect to the carrier concentrations are computed for the following reason. It can easily be seen that the derivatives of R^{SRH} with respect to n or p always increase the diagonal dominance in the linear system of equations, a welcome effect. Unfortunately, the derivatives of R^{AU} with respect to n or p, may decrease the diagonal dominance. Such a negative contribution to the main diagonal can possibly destroy the definiteness of the resulting linear system. Therefore the negative contributions to the main diagonal are discarded in Gummel's algorithm.

The discretization of the carrier continuity equations has to be done very carefully due to the layer–like behaviour of the variations of the carrier concentrations. In MINIMOS, the Scharfetter–Gummel interpolation scheme [17], modified for the carrier temperature dependent mobilities, is used. The box integration scheme assuming constant R within a box produces an nonsymmetric, but positive definite system of linear equations, again with property A.

For a vanishing derivative of R with respect to the carrier concentration n or p, respectively, only semidefiniteness of the equations is guaranteed, since each column sum of the offdiagonal elements equates the main diagonal element for all variables. Since we assume our problem to be well–posed, a zero eigenvalue and hence a singular system is unlikely to occur.

The solution of the nonsymmetric linear systems in the two-dimensional version of MINIMOS is performed by Gaussian elimination. This is due to the fact that the number of the mesh points is sufficiently small (for most simulations) to outperform any iterative solver. Moreover, there is the argument of the guaranteed stability of the LU decomposition. For the three-dimensional case this is no longer true. An optimized sparse matrix solver for the three-dimensional carrier would allocate roughly $(NX \leq NY \leq NZ)$

$$(NXY)^2 \cdot \frac{NZ-1}{2}$$

words, the corresponding flops would be greater than

$$NXYZ \cdot (14 + NXY\,(5 + 0.5 \cdot NXY))$$

These numbers are prohibitive for computers without very large primary and secondary storage space. Thus, iterative solution methods have to be used.

3 Symmetrization of the Coefficient Matrices

The coefficient matrices of the linear systems arising from the discretized transport equations are similar to symmetric positive definite matrices. This can be shown as follows. The nonsymmetry in the coefficient matrices is introduced by a different sign in the Bernoulli weights of two corresponding coefficients in the Scharfetter–Gummel [17] interpolation scheme. For two consecutive points in the naturally ordered scheme, e.g. the east coefficient of point i, $A(i)$, and the west coefficient of point $i+1$, $B(i+1)$, we have, for the electrons as carriers, ignoring equal factors for both coefficients:

$$\begin{aligned} A(i) &= \mathcal{B}\left(\tfrac{\psi_{i+1}-\psi_i}{U_t}\right) \\ B(i+1) &= \mathcal{B}\left(\tfrac{\psi_i-\psi_{i+1}}{U_t}\right) \end{aligned}$$

where $\mathcal{B}$ denotes the Bernoulli function

$$\mathcal{B}(x) = \frac{x}{exp(x)-1}$$

¿From these relations one can see immediately that the matrix $\bar{\mathbf{A}}$

$$\bar{\mathbf{A}} = \mathbf{W}^{-1} \cdot \mathbf{A} \cdot \mathbf{W}$$

with $W_{ii} = exp\left(\frac{\psi_i}{2 \cdot U_t}\right)$ is symmetric, hence $\mathbf{W}^{-1}$ is a *symmetrization* matrix. The definiteness of $\bar{\mathbf{A}}$ follows from the definiteness of **A**. For the hole transport equation the symmetrization matrix is **W**.

The problem with these matrices is their enormous dynamical range. The maximum allowed scaled absolute bias voltage at room temperature in MINIMOS is $\frac{20\ Volts}{U_t} = 775$. An algorithm using the symmetrization matrix would have to compute inner products of the form $\langle \mathbf{W}x, \mathbf{W}y \rangle$ and $\langle \mathbf{W}^{-1}x, \mathbf{W}^{-1}y \rangle$ over the previously sketched number range. Though there is the possibility to scale the symmetrization matrix appropriately, the unavoidable truncation error causes severe convergence problems on computers with only a medium exponent range. This strong dependence on the machine hardware has led us to the conclusion that the symmetrizability of the semiconductor transport equations is difficult to exploit and therefore only of restricted practical interest.

For scientific purposes we have implemented the symmetrized line `CJ-CG` [4] method and the `ILU` preconditionend `RF -CG` method (`ORTHORES`) acceleration. On the conference, we shall present numerical results, carried out in quad-precision on a VAX 8800 computer.

4 The NSPCG Package

A large number of algorithms for the solution of general nonsymmetric linear systems have been published in the last few years. In our application, all of these methods are, as far as we have investigated, applicable only with an effective preconditioner. An existing effective preconditioner brings many distinct methods to convergence in reasonable time and is thus crucial for the application of iterative accelerators to the transport equations.

A fine tool to study the applicability of various iterative schemes together with a number of established preconditioners on a certain application is the NSPCG package [12], included in the ITPACK2C package from the Center of Numerical Analysis at the university of Texas in Austin [9][8].

In the following section we present results obtained from the NSPCG package which was installed together with MINIMOS 5. The NSPCG allows both a large number of preconditioners together with many accelerators. Testing all combinations would be a very tedious task. Moreover, it has to be stated that a large variety of MINIMOS test problems have to be solved in a sufficient manner. Finding the best pair (preconditioner, accelerator) over a given set of model problems is a two-dimensional discrete optimization problem. Fortunately, it turns out that only a few combinations of those possibilities meet our requirements as will be outlined in the following.

5 Selecting an Efficient Preconditioner

At the beginning of our development, a block-Jacobi preconditioner was used, either as a line or as a plane preconditioner. This choice was motivated by our Poisson equation solver, the line `CJ-CG` method [5]. Unfortunately, this precondititioner was not stable for all our test examples. Especially at high bias voltages, the method failed and was sensitive to the selection of the starting vector. Moreover, the usual direction sensitivity due to the line or plane elimination process is most inconvenient and forces the swapping of the matrix to the most favourable direction, which was in most cases a plane orthogonal to the main current flow. We believe that the block Jacobi preconditioners are safely applicable only for low to medium biased device simulations.

Fortunately, a most favourable preconditioner was found in the incomplete LU factorization technique [7], denoted by `IC(k)`.

The index `k` denotes a controllable sparsity pattern along the matrix diagonals, `k=0` denoting no fill-in except the original matrix nonzero pattern, `k=1` allows fill-in caused by the original nonzero pattern but no further and so on. As expected, a higher degree of fill-in reduces the iteration count at the same time increasing the number of operations per iterations and the memory requirements. Obviously, the parameter k poses an optimization problem, namely choosing the appropriate degree of fill-in in order to minimize the flop count

per iteration. This problem is addressed in the next section.
One closely related preconditioner is the modified incomplete factorization denoted by MIC(k) [2]. Here the diagonal pivots of the factorization are adjusted such that **Q-A** has zero row sums, where **Q** is the inverse of the preconditioning matrix. We observed that none of the modified incomplete factorizations including the block versions behaved satisfactorily. However, choosing the diagonal pivots such that **Q-A** has zero ***column*** sums produced roughly the same iteration count as the unmodified incomplete factorization. Selecting a modification factor ω in the interval $[0,1]$ in order to tune the modification, we observed a decrease in the iteration count of up to 20 percent at $\omega = 0.8$.
There are a number of other preconditioners such as the SSOR, least squares polynomial, Neumann polynomial and their line and/or block variants. None of them is in our opinion competitive with the mere ILU factorization.

6 Choice of Accelerator

With ILU preconditioning of various levels a number of accelerators converges. We categorize the iterative methods together with the accelerator implementation as follows:

- Normal equations (LSQR)
- Generalized conjugate gradients (ORTHOMIN(1),ORTHORES(1),GMRES(1))
- Lanczos methods (LANMIN,LANRES,BCGS)

Simulations carried out with the above accelerators result in the following observations. LSQR [13] converges monotonically but with extremely low speed due to the bad condition number of the normal equations. For high bias voltage simulations, convergence stagnation was observed.
USYMLQ as well as its counterpart USYMQR[16], two methods working on a quasi Krylov space, converge monotonically with significantly higher speed than LSQR. ORTHOMIN(1) and ORTHORES(1), however, converge definitely fast. The iteration history of both is quite similar, reflecting their algorithmic closeness in the non-symmetric but definite case. The parameter 1 denotes the number of back vectors to be used in the purely truncated algorithm, for which a value of 5 was used in our numerical experiments.
It is interesting to notice that both ORTHODIR [23] [6] and LANDIR, its Lanczos equivalent, fail to converge.
The Lanczos methods represented by LANMIN and LANRES are altogether relatively similar in convergence and yield roughly the same iteration count as ORTHOMIN and ORTHORES.
GMRES(1), the general minimum residual method of Saad and Schultz [15], converges monotonically with slightly more iterations than the Lanczos methods but with significantly less CPU time per iteration. We suggest that the IC(1) preconditioned GMRES(5), providing the parameters for the stopping criterion automatically, is the only viable alternative to the IC(1) preconditioned BCGS [21]. The BCGS in its polynomial formulation computes the square of the corresponding error-reducing polynomial of LANMIN and therefore yields significantly faster convergence. In many cases, the speed is slightly less than twice the speed of LANMIN.
To demonstrate the performance as well as the impact of the various degrees of fill-in in the incomplete factorization an example is given below. An n–channel MOSFET simulation is taken with 3V drain and 1V gate bias. The hole carrier transport equation is selected from the first full Gummel iteration. The mesh is 33 x 33 x 33 in NX, NY and NZ direction. The standard stopping criterion

$$\left[\frac{\langle \tilde{z}^n, \tilde{z}^n \rangle}{\langle u^n, u^n \rangle}\right]^{\frac{1}{2}} < \zeta$$

with $\zeta = 10^{-6}$ was used, where $u^{(n)}$ denotes the current iterate and $\tilde{z}^n = Q_R^{-1} Q_L^{-1} r^n$ the current pseudoresidual. The maximum iteration count was set to 100 and a real workspace limit of $50 \cdot NXYZ$ was fixed. The table below shows the results.

The first column denotes the iteration count, the second the consumed CPU time in seconds on one processor of a VAX 8800:

METH	ORES		GMRES		LANMIN		LANRES		BCGS	
IC(0)	-	-	-	-	72	294	BR	BR	40	170
IC(1)	38	161	37	146	34	172	34	197	21	112
IC(2)	33	171	31	181	25	198	25	262	16	138
IC(3)	-	-	-	-	21	404	21	420	12	196

A hyphen denotes that the method, though obviously convergent, failed to converge within the iteration limit. For IC(0) all methods except BCGS and LANMIN failed to converge. A breakdown of LANRES is encountered at a relative accuracy of 10^{-5}. For IC(3) an excess of memory consumption took place for the first two methods. It can be seen that LANMIN and LANRES, both with the same iteration count, requires slightly less than twice the iteration count of BCGS.

Obviously, the BCGS method with IC(1) preconditioning is a clear winner. For a no-fill vector computer implementation, BCGS with IC(0) is recommended. BCGS

7 Implementations on Scalar, Parallel and Vector Computers

MINIMOS 5 as well as its solver package is being developed on a VAX 8800 with two processors. Independently from the NSPCG package, FORTRAN production code for the point/line/plane Jacobi preconditioner, the IC(0), MIC(0), and their line variants were developed. As accelerators we implemented the symmetrized CG, LANRES and BCGS. We stress that the iterative results are basically the same as the NSPCG parameters. Execution speed of our production code was slightly less than doubled.

A two-processor implementation of the preconditioned Lanczos algorithms was installed on our VAX. Such a parallelization is relatively simple, because the Lanczos Methods use both the original and the transposed system. A speedup of almost hundred percent was achieved.

In cooperation with Siemens-PDS in Vienna, the IC(0) preconditioned BCGS method for the carrier continuity equations and the line CJ-CG method for the Poisson equation was recently installed and optimized on the Fujitsu VP200 supercomputer at Siemens Munich.

More detailed implementational notes including our implementation of the hyperplane ordering method [22] for the computation of the IC(0) factorization will be made at the confere nce, together with measuring data from the installation on the CRAY-2 supercomputer at the RUS Stuttgart.

References

[1] Gummel, H.K., "A Selfconsistent Iterative Scheme for Onedimensional Steady State Transistor Calculations", *IEEE ED-11*, pp. 455-465, 1964.

[2] Gustafsson, I. *Stability and Rate of Convergence of Modified Incomplete Cholesky Factorization Methods.* Doctoral dissertation. Chalmers University of Technology and the University of Göteborg, April 1979.

[3] Hänsch, W., Selberherr, S. "MINIMOS 3: A MOSFET Simulator that Includes Energy Balance", *IEEE ED-34*, pp.1074-1078, 1987.

[4] Hageman, L., Luk, F, Young, D.M. "On the Equivalence of Certain Iterative Methods" *SIAM Journal of Numerical Analysis,* Vol. 17, No. 6, Dec. 1980, pp. 852-973.

[5] Hageman, L. and Young, D.M. *Applied Iterative Methods.* New York: Academic Press, Inc., 1981.

[6] Jea, K.C. and Young, D.M. "On the Simplification of Generalized Conjugate Gradient Methods for Nonsymmetrizable Linear Systems." ***Linear Algebra and its Applications,*** Vol 52/53, 1983, pp. 399-417.

[7] Kershaw, D.S. "The Incomplete Cholesky–Conjugate Gradient Method for the Iterative Solution of Systems of Linear Equations." ***Journal of Computational Physics,*** Vol. 26, pp. 43-65.

[8] Kincaid, D., Oppe, T., Respess, J., and Young, D. "ITPACKV 2C User's Guide." CNA-191, Center for Numerical Analysis, University of Texas, Austin, Texas, 78712, February 1984.

[9] Kincaid, D., Respess, J., Young, D., and Grimes, R. "Algorithm 586 ITPACK 2C: A FORTRAN Package for Solving Large Sparse Linear Systems by Adaptive Accelerated Iterative Methods." *ACM Transactions on Mathematical Software,* Vol. 8, No. 3, September 1982, pp. 302-322.

[10] Lindorfer, P., Selberherr S. "MESFET Analysis with MINIMOS",
Proc. ESSDERC '89 Conf., 1989.

[11] Nylander, J.O., Masszi, F., Selberherr, S., Berg, S. "Computer Simulations of Schottky Contacts with a Non-Constant Recombination Velocity",
Solid State Electronics, Vol. 32, No. 5, pp.363-367, 1989.

[12] Oppe, T.C., Joubert, W.D., and Kincaid, D.R. "NSPCG User's Guide." Center of Numerical Analysis, The University of Texas at Austin.

[13] Paige, C.C. and Saunders, M.A. "LSQR: An Algorithm for Sparse Linear Equations and Sparse Least Squares." *ACM Transactions on Mathematical Software,* Vol. 8, No. 1, March 1982, pp. 43-71.

[14] Saunders, M.A., Simon, H.D., Yip, E.L. "Two Conjugate–Gradient–Type Methods for Unsymmetric Linear Equations." *SIAM Journal of Numerical Analysis,* Vol. 25, No. 4, August 1988, pp. 927-640.

[15] Saad, Y. and Schultz, M.H. "GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems." *SIAM Journal of Scientific and Statistical Computing,* Vol. 7, No. 3, July 1986, pp. 856-869.

[16] Saunders, M.A., Simon, H.D., Yip, E.L. "Two Conjugate–Gradient–Type Methods for Unsymmetric Linear Equations." *SIAM Journal of Numerical Analysis,* Vol. 25, No. 4, August 1988, pp. 927-640.

[17] Scharfetter, D.L., Gummel, H.K., "Large–Signal Analysis of a Silicon Read Diode Oscillator." *IEEE ED-16*, pp. 64-77, 1969.

[18] Selberherr, S. "Analysis and Simulation of Semiconductor Devices",
Springer-Verlag Wien New York, ISBN 3-211-81800-6, 1984.

[19] Selberherr, S. "MOS Device Modeling at 77K",
IEEE ED-36, pp.1464-1474, 1989.

[20] Selberherr, S. "The Status of MINIMOS",
in: *Simulation of Semiconductor Devices and Processes,*
edited by: K.Board, D.R.J.Owen,
ISBN 0-906674-59-X, pp.2-15, 1986.

[21] Sonneveld, P. "CGS, A fast Lanczos–Type Solver for Nonsymmetric Systems" *SIAM Journal of Sci. Stat. Computing* Vol. 10, No. 1, Jan. 1989, pp. 36-52.

[22] Vorst, H. "High Performance Preconditioning" *SIAM Journal of Sci. Stat. Computing* Vol. 10, No. 6, Nov. 1989, pp. 1174-1185.

[23] Young, D.M. and Jea, K.C. "Generalized Conjugate Gradient Acceleration of Nonsymmetrizable Iterative Methods." *Linear Algebra and its Applications,* 34:159-194 (1980).