VISTA Status Report December 2002

H. Ceric, K. Dragosits, A. Gehring, S. Smirnov,
V. Palankovski, S. Selberherr

Institute for Microelectronics
Technische Universität Wien
Gusshausstrasse 27-29
A-1040 Vienna, Austria

Contents

1 An Adaptive Grid Approach for the Simulation of Electromigration **1**

1.1 Introduction 1

1.2 Theoretical Formulation 1

1.3 Applied Diffuse Interface Model 2

1.4 Numerical Implementation 3

1.5 Setting of the Initial Order Parameter Profile and Initial Grid Refinement $\Lambda_h(0)$ 3

1.5.1 Finite Element Scheme 3

1.5.2 Maintaining the Grid During Simulation 4

1.5.3 Grid Adaptation 4

1.5.4 Solving Procedure 6

1.6 Results 7

1.7 Conclusion 10

2 2D Mechanical Quantum Effects in Ultra-Short CMOS Devices **11**

2.1 Introduction 11

2.2 The QM Model 11

2.2.1 Approximation of the Wave Function 11

2.2.2 Approximation of the Energy Band Structure 12

2.3 Simulation Results 13

2.3.1 Simulation of MOS Capacitors 13

2.3.2 Simulation of a CMOS Transistor 13

2.4 Outlook 13

3 A Tunneling Model Based on a Non-Maxwellian Energy Distribution **16**

3.1 Introduction 16

3.2 Tunneling Model 16

3.3 Distribution Function 17

3.4 Results 19

3.5 Conclusion 20

4 Investigation of the Electron Mobility in Strained $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ 22

4.1 Introduction 22

4.2 Strain Effects. Undoped Material 22

4.3 Inclusion of Ionized Impurity Scattering Rate 23

4.4 Results 24

4.5 Conclusion 24

5 Effects of Stress-Induced Bandgap Narrowing 29

5.1 Introduction 29

5.2 Experiments 29

5.3 Results and Discussion 30

5.4 Conclusions 31

1 An Adaptive Grid Approach for the Simulation of Electromigration

For tracking electromigration induced evolution of voids a diffuse interface model is applied. We assume an interconnect as two-dimensional electrically conducting via which contains initially a circular void. The diffuse interface governing equation was solved applying a finite element scheme with a robust local grid adaptation algorithm. Simulations were carried out for voids exposed to high current. An influence of the void dynamics on the resistance of interconnect is investigated. In the case of the interconnect via it was shown that a migrating void exactly follows the current flow, retaining its stability, but due to change of shape and position causes significant fluctuations in interconnect resistance.

1.1 Introduction

Failure of metallic interconnects in integrated circuits caused by electromigration has long been a key reliability concern which is further accentuated by the continuing trend of miniaturization. The phenomenon of electromigration is a mechanism for transport of matter by high electric current densities which produce damage in the interconnect lines. Failure results either from voids growing over the entire line width or extrusions which cause short circuits to neighboring lines.
Modeling the micromechanics of void evolution is a long-standing scientific problem. It began with sharp interface models requiring an explicit finite element tracking of void surfaces during the course of evolution [1]. Later, prompted by the complexity of void surfaces, diffuse interface models were introduced [2]. Diffuse interface models circumvent computationally costly explicit surface tracking by application of a smooth order parameter field for representation of void structures. An alternative diffuse interface model based on the double obstacle potential was proposed in [3] where the computation is simplified by reduction of order parameter profiles evaluation only to the void-metal interfacial area.
The main disadvantages of these diffuse interface models [3, 4] is their requirement of structured underlaying grids for the order parameter evaluation and also their restricted facility to reach higher resolution of an order parameter profile in the void-metal interfacial area. To overcome these drawbacks we solve the diffuse interface model governing equation with a finite element scheme coupled with a powerful grid adaptation algorithm. The robustness of the developed finite element approach with respect to the underlying grid structure makes it possible to efficiently simulate the damage induced by electromigration in complex interconnect geometries.

1.2 Theoretical Formulation

We assumed unpassivated monocrystal isotropic interconnects where stress phenomena can be neglected. An interconnect is idealized as two-dimensional electrically conducting via which contains initially a circular void. For simplicity we also neglect the effects of grain boundaries and lattice diffusion. In this case there are two main forces which influence the shape of the evolving void interface: the chemical potential gradient and electron wind. The first force causes self-diffusion of metal atoms on the void interface and tends to minimize energy which results in circular void shapes. The electron wind force produces asymmetry in the void shape depending on the electrical field gradients.
Including contributions from both electromigration and chemical potential-driven surface diffusion gives the total surface atomic flux, $\mathbf{J_A} = J_A\mathbf{t}$, where $\mathbf{t}$ is the unit vector tangent to the void surface [1, 5]

$$J_A = D_s(-|e|Z^* E_s + \gamma_s \Omega\, \mathbf{t} \cdot \nabla_\mathbf{s} \kappa) \tag{1}$$

Z^* is the effective valence, e is the charge of an electron, $E_s \equiv \mathbf{E_s} \cdot \mathbf{t}$ is the local component of the electric field tangential to the void surface, κ is the local surface curvature, and ∇_s is the surface gradient operator;

$\kappa \equiv \nabla \cdot \mathbf{n}$, where $\mathbf{n}$ is the local unit vector normal to the void surface. Further, γ_s is the surface energy, Ω is the volume of an atom, and D_s is given trough Arrhenius' law:

$$D_s = \frac{D_0 \delta_s}{k_B T} exp\Big(\frac{-Q_s}{k_B T}\Big) \tag{2}$$

Here, δ_s is the thickness of the diffusion layer, k_B is Boltzmann's constant, T is the temperature, Q_s is the activation energy for the surface diffusion, and D_0 is the pre-exponential for mass diffusion. Equation (1) is the Nernst-Einstein equation, where the sum in the parentheses on the right side expresses the driving force. Mass conservation gives the void propagation velocity normal to the void surface, v_n, through the continuity equation [6, 7, 5],

$$v_n = -\Omega \cdot \nabla_s \mathbf{J_A}. \tag{3}$$

The electric field $\mathbf{E}$ in the interconnect is irrotational, $\nabla \times \mathbf{E} = \mathbf{0}$ and it can be written as the gradient of the electric potential, i.e.,

$$\mathbf{E} \equiv -\nabla \mathbf{V}. \tag{4}$$

The field is also solenoidal, i.e.,

$$\nabla \cdot \mathbf{E} = \mathbf{0}. \tag{5}$$

Equations (4) and (5) imply that the potential V obeys Laplace's equation,

$$\Delta V = 0 \tag{6}$$

The void surface and the interconnect's edges are modeled as insulating boundaries, i.e., $\nabla V \cdot \mathbf{n} = 0$ at these boundaries.

1.3 Applied Diffuse Interface Model

Direct numerical implementation of equations (1)-(6) demands explicit tracking of the moving void-metal interface. The interface is described by specifying a large number of points on it. Over the time the void-metal interface evolves and changes morphology and even more points may be required to accurately describe it. Such techniques are very complicated to implement and also tend to have rather poor numerical stability.

In the diffuse interface models void and metal area are presented through the order parameter ϕ which takes values $+1$ in the metal area, -1 in the void area and $-1 < \phi < +1$ in the void-metal interface area. Demanding explicit tracking of the void-metal interface is not needed and models do not require boundary conditions to be enforced at the moving boundary. Diffuse interface models are, as we will see in the next sections, quite simple to implement numerically.

The model equations for the void evolving in an unpassivated interconnect line are the balance equation for the order parameter ϕ [2, 8, 4, 3],

$$\frac{\partial \phi}{\partial t} = \frac{2D_s}{\epsilon \pi} \nabla \cdot (\nabla \mu - |e| Z^* \nabla V) \tag{7}$$

$$\mu = \frac{4 \Omega \gamma_s}{\epsilon \pi} (f'(\phi) - \epsilon^2 \Delta \phi) \tag{8}$$

and for the electrical field

$$\nabla \cdot (\sigma(\phi) \nabla V) = 0 \tag{9}$$

where μ is the chemical potential, $f(\phi)$ is the double obstacle potential as defined in [9, 10], and ϵ is a parameter controlling the void-metal interface width. The electrical conductivity was taken to vary linearly from the metal ($\sigma = \sigma_{metal}$) to the void area ($\sigma = 0$) by setting $\sigma = \sigma_{metal}(1 + \phi)/2$. The equations (7)

and (9) are solved on the two-dimensional polygonal interconnect area T.
It has been proven [8, 3] that in the asymptotic limit for $\epsilon \to 0$ the diffuse interface model defined by equations (7)-(9) describes the same voids-metal interface behavior like equations (1)-(6). The width of the void-metal diffuse interface is approximately $\epsilon\pi/2$, and in order to numerically handle sufficiently thin interfaces one needs a very fine locally placed grid around it.

1.4 Numerical Implementation

The equations (7)-(9) are solved by means of the finite element method on the sequence of the grids $\Lambda_h(t_0 = 0), \Lambda_h(t_1), \Lambda_h(t_2)$ each one adapted to the position of the void-metal interface from the previous time step. The initial grid $\Lambda_h(0)$ is produced by refinement of some basic triangulation T_h of area T according to the initial profile of order parameter ϕ. The motivation of grid adaptation is to construct and maintain a fine triangulated belt of width $\epsilon\pi$ in the interconnect area where $-1 < \phi < +1$, respectively, where the void-metal interface area is placed.

1.5 Setting of the Initial Order Parameter Profile and Initial Grid Refinement $\Lambda_h(0)$

The initial order parameter profile depends on the initial shape of the void $\Gamma(0)$ and can be expressed as

$$\phi(x,y) = \begin{cases} +1 & \text{if} \quad d > \frac{\epsilon\pi}{2}, \\ \sin(\frac{d}{\epsilon}) & \text{if} \quad |d| \leq \frac{\epsilon\pi}{2}, \\ -1 & \text{if} \quad d < -\frac{\epsilon\pi}{2} \end{cases} \tag{10}$$

Where $d = dist(P(x,y), \Gamma(0))$ is the signed normal distance of the point from the initial interface $\Gamma(0)$. To obtain sufficient resolution of this initial profile, the basic grid T_h is transformed into grid $\Lambda_h(0)$ obeying the following *initial grid refinement criterion (IGRC)* for the circular void with center O and radius r:

$$|dist(C_K, O) - r| \leq \frac{\epsilon\pi}{2} \quad \wedge \quad h_K > \frac{\epsilon\pi}{n} \tag{11}$$

n is the chosen number of grid elements across the void-metal interface width, h_K is the longest vertex of the triangle K, and C_K is its center of gravity. Now an adaptive algorithm defined in Section 1.5.3 transforms the basic grid T_h into initial grid $\Lambda_h(0)$ according to $IGRC$.

1.5.1 Finite Element Scheme

Let $\Lambda_h(t_n)$ be a triangulation of the area T at the discrete time t_n, and let $\{\phi_i^{n-1}\}_{i=0}^{N-1}$ be discrete nodal values of the order parameter on this triangulation.
A finite element based iteration for solving (7) on grid $\Lambda_h(t_n)$ and evaluating the order parameter at the time $t_{n+1} = t_n + \Delta t$ consists of two steps [11]:
Step 1. For the kth iteration of $n+1^{th}$ time step the linear system of equations has to be solved:

$$\epsilon\frac{\pi}{2}M_{ii}\phi_i^{n+1,k} + \Delta t D_s K_{ii}\mu_i^{n+1,k} = \alpha_i \tag{12}$$

$$M_{ii}\mu_i^{n+1,k} - \tau\left(\frac{1}{\epsilon}M_{ii} + \epsilon K_{ii}\right)\phi_i^{n+1,k} = \beta_i, \tag{13}$$

where

$$\alpha_i = \epsilon\frac{\pi}{2}M_{ii}\phi_i^n - \Delta t D_s \sum_{i \neq j} K_{ij}\mu_j^{n+1,k-1} \tag{14}$$

$$\beta_i = \tau\epsilon \sum_{i\neq j} K_{ij}\phi_j^{n+1,k-1} - |e|Z^* M_{ii} V_i^n \tag{15}$$

for each $i = 0, N-1$ of the nodal values $(\phi_i^{n+1}, \mu_i^{n+1})$ of the triangulation $\Lambda_h(t_n)$. $[M]_{ij}$ and $[K]_{ij}$ are the lumped mass and stiffness matrix, respectively and $\tau = \frac{4\Omega\gamma_s}{\pi}$.
Step 2. All nodal values $\{\phi_i^{n+1}\}_{i=0}^{N-1}$ are projected on $[-1, 1]$ by a function

$$\rho(x) = max(-1, min(1, x)). \tag{16}$$

For solving (9) a conventional finite element scheme is applied [12].

1.5.2 Maintaining the Grid During Simulation

After an order parameter was evaluated on the $\Lambda_h(t_n)$ a grid needs to be adapted according to the new void-metal interface position. Therefore it is necessary to extract all elements which are cut by the void-metal interface in grid $\Lambda_h(t_n)$. The following condition is used: Let us take a triangle $K \in \Lambda_h(t_n)$ and denote its vertices as P_0, P_1, P_2. The triangle K belongs to the interfacial elements if for the values of the order parameter ϕ at the triangle's vertices holds $\phi(P_0)\phi(P_1) < 0$ or $\phi(P_1)\phi(P_2) < 0$. We assume that an interface intersects each edge of the element only once. The set of all interfacial elements at the time t_n is denoted as $E(t_n)$. The centers of gravity of each triangle from the $E(t_n)$ build the interface point list $L(t_n)$. The distance of the arbitrary point Q from $L(t_n)$ is defined as

$$dist(Q, L(t_n)) = \min_{P\in L(t_n)} dist(Q, P). \tag{17}$$

Thus we can define the *transitional grid refinement criterion TGRC*

$$dist(C_K, L(t_n)) \leq \frac{\epsilon\pi}{2} \quad \wedge \quad h_K > \frac{\epsilon\pi}{n} \tag{18}$$

The grid adaption for the next time step evaluation of the order parameter ϕ is now completed with respect to $TGRC$ by as defined in the next section.

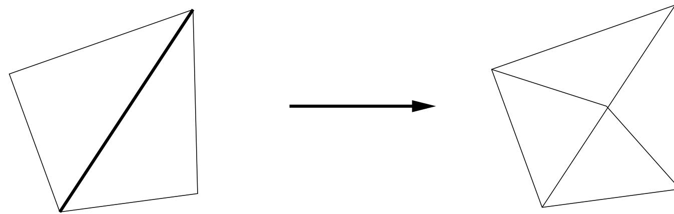

Figure 1: Atomic refinement operation.

1.5.3 Grid Adaptation

The grid adaption algorithm used in this work is a version of the algorithm introduced in [13] and consists of a *grid refinement algorithm* and a *grid coarsening algorithm*. The refinement algorithm is based on recursive bisecting of triangles. A triangle is marked for refinement if it complies with some specific refinement criterion $COND$. In our application we used $COND = IGRC$ for the initial refinement and $COND = TGRC$ for the grid maintaining refinement. For every triangle of the grid, the longest one of its edges is marked as *refinement edge*. The element and its neighbor element which also contains the same refinement edge are refined by bisecting this edge (Fig. 1). We can define refinement of the element in the following way:

Algorithm 1. *refinement of the element*

```
subroutine
recursive_refine(element)
{
 if neighbor has no compatible
ref_edge
 recursive_refine(neighbor)
 bisect(element)
 bisect(neighbor)
}
```

Now, the overall refinement algorithm can be formulated as follows:

Algorithm 2. *refinement of the grid*

```
subroutine refine_grid(COND)
{
 for n ≤ max_refinement_depth
 {
  for all elements
  if (COND) mark element for
refinement
  if no element is marked
  break refinement loop
  for all elements
  recursive_refine(element)
 }
}
```

The parameter `max_refinement_depth` limits the number of bisecting of each triangles of the grid. The grid is refined until there is no more element marked for refinement or the maximal refinement depth is reached.
The coarsening algorithm is more or less the inverse of the refinement algorithm. Each element that does not fullfil criterion $COND$ is marked for coarsening. The basic idea is to find the father element whose refinement produced the element in consideration.

Algorithm 3. *coarsening of the element*

```
subroutine coarsen(element)
{
 get the element_father
 get the father_neighbor on
ref_edge
 coarse the element_father
 coarse the father_neighbor
}
```

The following routine coarsens as many elements as possible, even more than once if allowed:

Algorithm 4. *coarsening of the grid*

```
subroutine coarsen_grid(COND)
{
 for all elements
 if (!COND) mark element for
coarsening
 do
 {
  do_coarsen_once_more = false
  for all elements
  if element is marked for
coarsening
  do_coarsen_once_more |=
coarsen(element)
 } until do_coarsen_once_more
is false
}
```

The complete adaption of the grid is reached by sequential invoking of the grid refinement and of the grid coarsening algorithm.

1.5.4 Solving Procedure

Combining all the components described in the Sections 1.5-1.5.3 we obtain the complete void evolution solving procedure (Fig. 2).

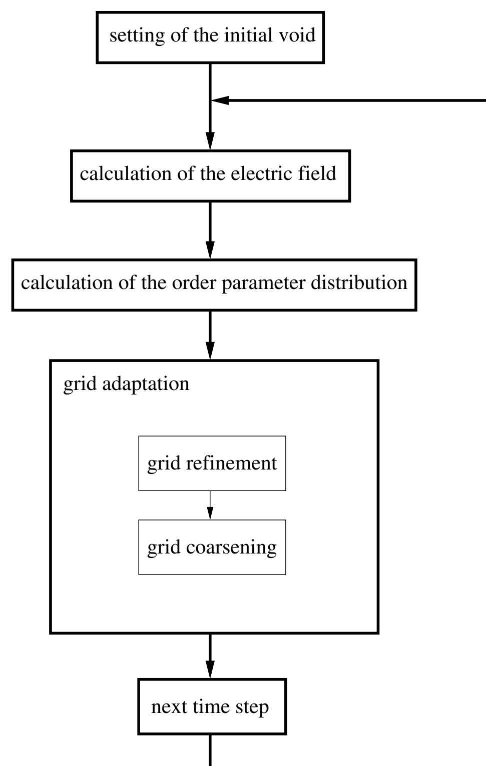

Figure 2: Void evolution solving procedure.

1.6 Results

In all simulations a circle was chosen as initial void shape. The resolution of the parameter ϕ profile can be manipulated by setting parameter n which is the mean number of triangles across the void-metal interface. On Fig. 3 initial grids for $n = 1$ and $n = 5$ are presented. We consider a two-dimensional, stress free,

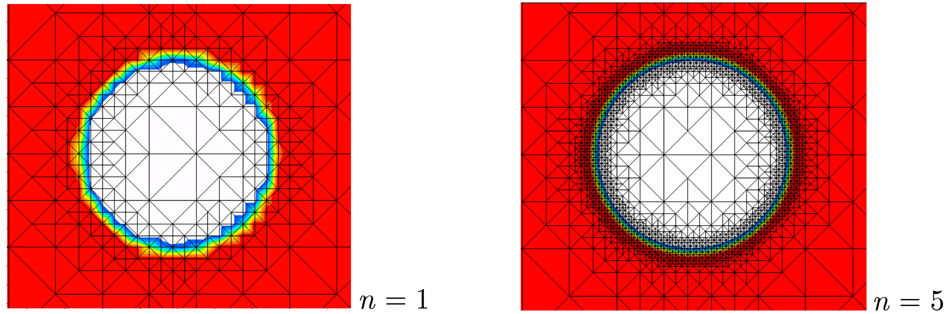

Figure 3: Initial grid refinements.

electricaly conducting interconnect via.

A constant voltage is applied between points A and B (Fig. 5). At B a refractory layer is assumed. Because of geometrical reasons there is current crowding in the adjacencies of the corners C and D. The analytical solution of equation (9) has at these points actually a singularity [12].
High electrical field gradients in the area around the corner points increase overall error of finite element scheme for the equation (9) which is overcome by applying an additional refinement of the basic mesh T_h according to the local value of the electric field gradient (Fig. 4). A fine triangulated belt area which is

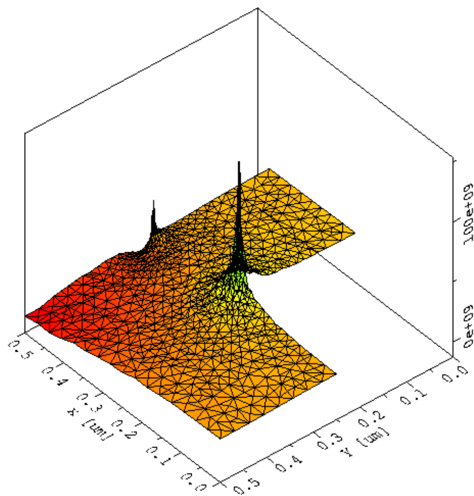

Figure 4: Profile of the current density at the corners of the interconnect (in A/m^2).

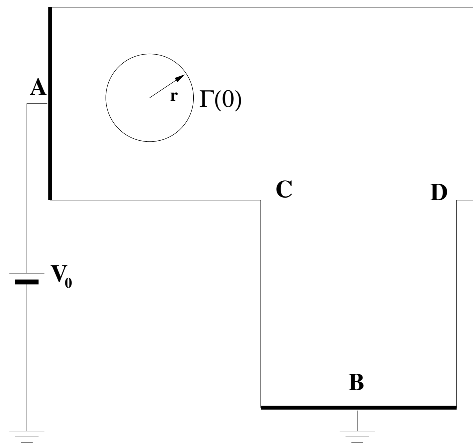

Figure 5: Interconnect via with initial void.

attached to the void-metal interface at the initial simulations step follows the interfacial area throughout the simulations whereby the interconnect area outside the interface is coarsened to the level of the basic grid T_h (Fig. 6). In our simulations a void evolving through straight part of the interconnect geometry exhibits similar shape changes as observed in the earlier models [1, 3]. There is also no significant fluctuation of the resistance during this period of interconnect evolution. The situation changes when the void evolves in the proximity of the interconnect corner. Due to current crowding in this area the influence of the electromigration force on the material transport on the void surface is more pronounced than the chemical

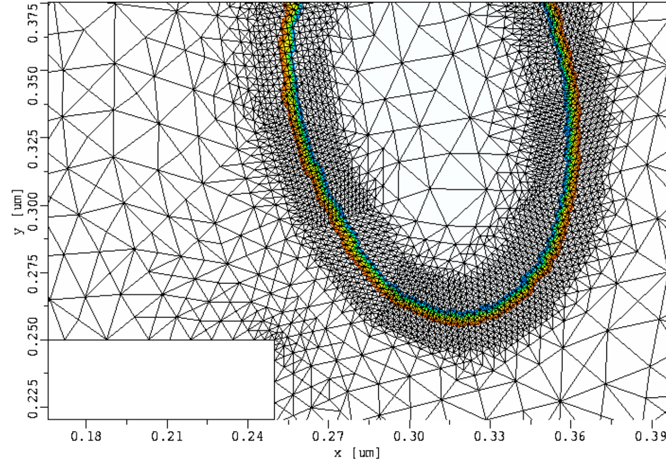

Figure 6: Refined grid around the void in the proximity of the interconnect corner.

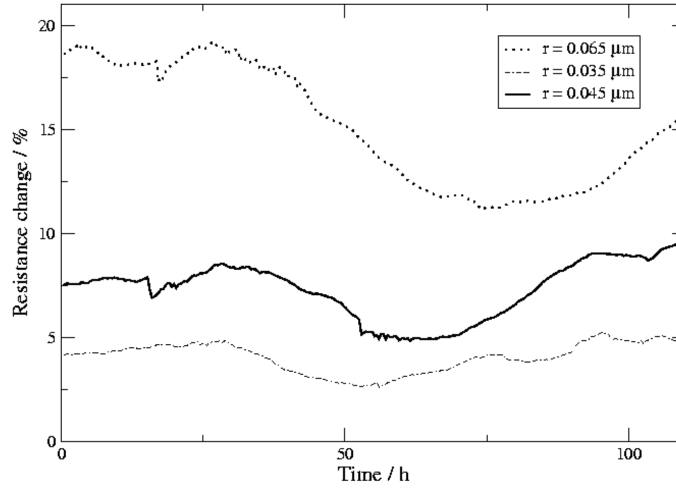

Figure 7: Time dependent resistance change during void evolution for the different initial void radius r.

potential gradient. This unbalance leads to higher asymmetry in the void shape then observed in the straight part of the interconnect. A void evolving in the proximity of the interconnect corner causes significant fluctuations in the interconnect resistance due to void asymmetry and position. The resistance change shows a charasteristical profile with the two peaks and a valley (Fig. 7). The extremes are more pronounced for the larger initial voids.
The capability of the applied adaptation scheme is also presented in the simulation of void collision with the interconnect refractory layer (Fig. 8).
The time step Δt for the numerical scheme (12)-(15) is fitted at the simulations begin taking into account inverse proportionality of the speed of the evolving void-metal interface to the initial void radius [1]:

$$\Delta t = \frac{\epsilon \pi r l}{2 D_s |e| Z^* V_0} \tag{19}$$

l is the characteristical length of via geometry. Appropriate choice of the time step ensures that the

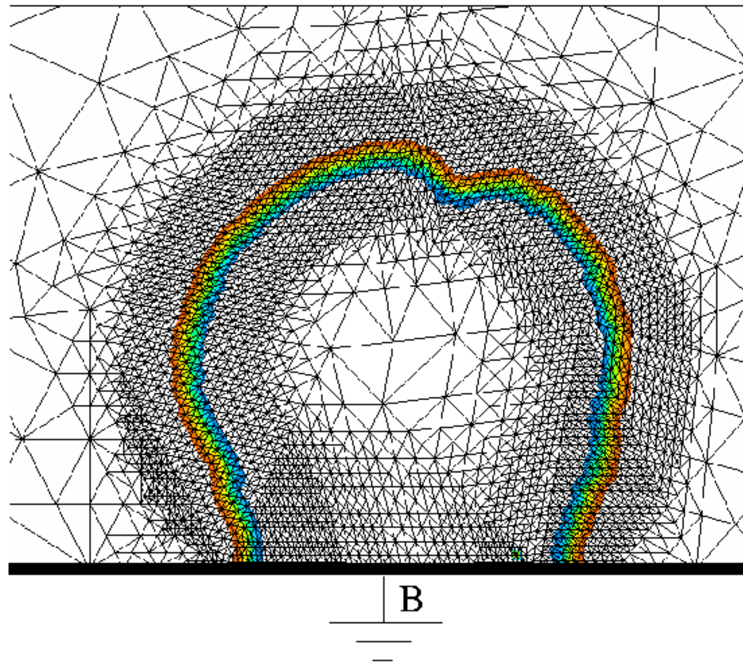

Figure 8: Grid adaptation in the case of void collision with the refractory layer.

evolving void-metal interface will stay inside the fine grid belt during the simulation. The dynamics of the evolving void-metal interface simulated with a the presented numerical scheme complies with the mass conservation law, the void area (where $\phi = -1$) remains approximately the same during the whole simulation. Notable area deviations during the simulation appear only if a relatively large factor ϵ has been chosen. As scaling length we took $l = 1\ \mu m$ and for the initial void radius $r_0 = 0.035\ l$, $r_1 = 0.045\ l$, and $r_2 = 0.065\ l$. Our simulations have shown that for all considered initial void radii, voids follow the electric current direction (Fig. 9) and do not transform in slit or wedge like formations which have been found to be a main cause for a complete interconnect failure [14]. Already with $\epsilon = 0.003\ l$ good approximations are achieved. The number of elements on the cross section of the void-metal interface was chosen between 6 and 10 with the interface width of $0.0015\ l\pi$.

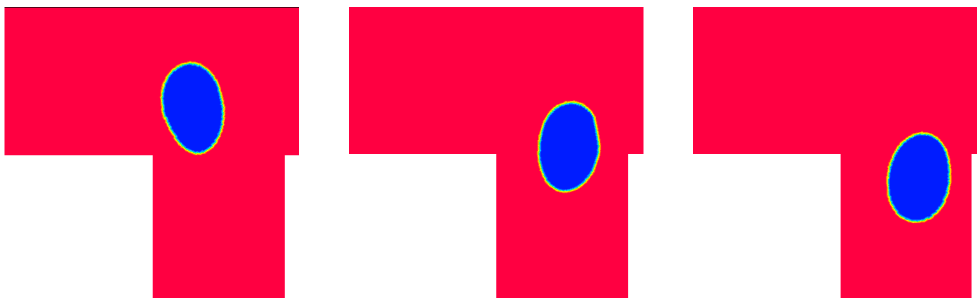

Figure 9: Void evolving through interconnect in the electric current direction

1.7 Conclusion

A governing diffuse interface equation for the order parameter coupled with the Laplace equation for the electrical field is solved using the finite element scheme. A dynamically adapted grid is maintained by a refinement-coarsening algorithm controlled by position, curvature, and width of the simulated void-metal interface which distributes the grid density in such a way that it allows an efficient simulation of evolving voids through large portions of a complex interconnect geometry. Due to high electrical current gradients in the proximity of the interconnect corners and overall asymmetry of the electrical field, voids exhibit specific faceting which was not observed in the case of straight interconnect geometries. The presented method is well suited for long time prediction of resistance change due to electromigration during the interconnect life time. The applied diffuse interface model extends readily to incorporate additional physical phenomena such as anisotropy, temperature variations, and bulk and grain boundary diffusion.

2 2D Mechanical Quantum Effects in Ultra-Short CMOS Devices

Quantum mechanical analysis of the quantum confinement of ultrashort CMOS is numerically very expensive. In this paper we present a macroscopic model, which includes a new approach to match the vertical carrier profile and combines it with a classical model in lateral direction. The simulation results show a significant improvement concerning the accuracy of the carrier profile and the C/V characteristics.

2.1 Introduction

It is well known that in consequence of the ever shrinking feature sizes quantum mechanical (QM) effects are getting more important for the performance of state-of-the-art CMOS devices. This issue is widely investigated and understood at physical level, but a full two-dimensional treatment is far from trivial and numerically very expensive.

Up-to-date macroscopic models incorporating QM effects were basically designed to fit the C/V characteristics of the investigated MOS structures [15][16][17] but not for a fully two-dimensional treatment. In this paper a new approach to model the quantum confinement near the channel surface is introduced which does not only match the C/V characteristics but also allows the two-dimensional analysis of advanced CMOS and similar devices. Furthermore the new model is capable to match the vertical carrier profile very accurately, thus offering new insight into the the actual properties of modern devices.

2.2 The QM Model

Classical device modeling leads to two significant inaccuracies concerning the carrier concentration near the channel surface.

First, the splitting of the conduction band into several discrete eigenvalues is not considered. This leads to an overestimation of the surface charge, as the energy difference between these discrete eigenvalues and the Fermi-level is bigger than the one from the bottom of the conduction band to the Fermi-level. Second, classical models do not consider that the shape of the wave functions reduces the carrier concentration near the surface as well. Consequently, a rigorous approach to simulate the carrier concentration has to take care of both effects, by offering approximations for the wave function and the actual band structure.

2.2.1 Approximation of the Wave Function

We model the first of these effects by a reduction of the density of states N_C near the interface applying an exponential shape function. This follows an approach proposed by Hänsch et al. [15],

$$N_\mathrm{C}(z) = N_\mathrm{C} \cdot \left(1 - e^{-(z+z_0)^2/\lambda_\mathrm{TH}^2}\right), \tag{20}$$

where z is the distance to the interface and z_0 is an offset to match the nonzero carrier concentration near the surface stemming from the finite barrier height. λ_TH is the thermal wavelength responsible for the reduction of the QM effect with increasing distance from the interface,

$$\lambda_\mathrm{TH} = \frac{\sqrt{2mkT}}{\hbar}. \tag{21}$$

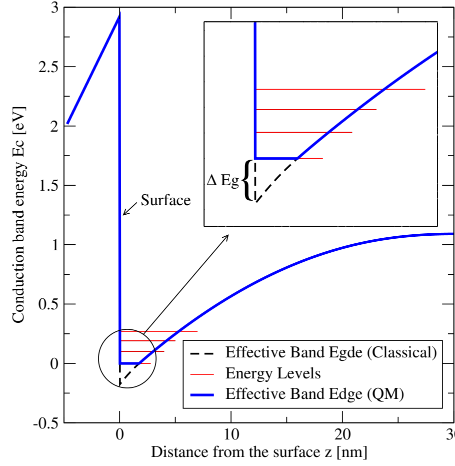

Figure 10: Band structure near the surface and comparison between the conduction band edge of the classical and the new approach

If this correction is applied the qualitative carrier distribution near the interface in strong inversion is reproduced quite well, but without consideration of band structure effects, this is not the case in the threshold voltage region [17].

2.2.2 Approximation of the Energy Band Structure

Fig. 10 shows the actual band structure near the surface of a MOS capacitor, calculated with a self-consistent Schrödinger-Poisson solver working with the effective mass approach [18]. Near the surface the lowest eigenenergy is significancy higher than the band edge, thus causing an overestimation of the charge when the classical simulation approach is applied. The basic idea of our model is to replace the effective band edge by the first discrete energy level (see Fig. 10). This seems reasonable as quantum mechanical calculations show that usually over 90% of the carriers are in this energy band.

We set the band edge at the surface to

$$E_{\mathrm{g,Surface}}^{\mathrm{QM}} = E_{\mathrm{g,Surface}}^{\mathrm{Classical}} + \Delta E_{\mathrm{g}}, \tag{22}$$

whereas $E_{\mathrm{g,Surface}}^{\mathrm{QM}}$ is the modified bandgap energy which is used in the Boltzmann statistics, $E_{\mathrm{g}}^{\mathrm{Classical}}$ is the bandgap according to the material specification, and ΔE_{g} is the applied correction. Our model pins the band edge $E_{\mathrm{g}}^{\mathrm{QM}}(z)$ inside the device to the value of $E_{\mathrm{g,Surface}}^{\mathrm{QM}}$ as long as $E_{\mathrm{g,Surface}}^{\mathrm{QM}} > E_{\mathrm{g}}^{\mathrm{Classical}}(z)$.

As the exact calculation of the first energy level is numerically expensive and requires the solution of the Schrödinger equation an approximation is used: The offset ΔE_{g} is approximated following a formulation of Van Dort et al. [17], which reads as

$$\Delta E_{\mathrm{g}} = \frac{13}{9} \cdot \beta \cdot \left(\frac{\epsilon}{4qkT} \right)^{1/3} |E_{\mathrm{surf}}|^{2/3}, \tag{23}$$

whereas $|E_{\mathrm{surf}}|$ is the magnitude of the electric field at the interface and ϵ is the permittivity of the semiconductor. $\beta = 4.1 \times 10^{-8}\mathrm{eVcm}$ is an empirical constant.

2.3 Simulation Results

The model was implemented into the device simulator MINIMOS-NT [19] and several typical structures were simulated in order to explore its one- and two-dimensional capabilities.

2.3.1 Simulation of MOS Capacitors

The one-dimensional capabilities of the new model were checked by simulations of a MOS capacitor. As reference the same structure was simulated using the self-consistent Schrödinger-Poisson solver. Fig. 11 shows a comparison of the C/V characteristics obtained with the new model, the Schrödinger-Poisson solver, and the Hänsch model. An excellent fit between the results from our new model and the quantum mechanical calculations is obtained, especially the onset of inversion is predicted very accurately. The overestimation of the capacity in strong inversion is a well known effect stemming from the different distribution statistics applied in the Schrödinger solver (Fermi-Dirac) and the device simulator (Boltzmann)(see e.g. [20]).

Fig. 12 shows a comparison of electron concentration profiles. It can be clearly seen that the new model offers a decent fit to the quantum mechanical results. The Hänsch model concentrates the charge closer to the surface, which is a consequence of the comparatively smaller band gap.

2.3.2 Simulation of a CMOS Transistor

The ability to reproduce the channel profile opens the door to a precise simulation of state-of-the-art CMOS devices. Fig. 13 shows the results obtained for a device with 50nm gate length using the QM approximation and compares them to a classical simulation. The influence of the quantum confinement, which results in a reduction of the drain current, is clearly visible.

It is remarkable that by introducing our model into the simulator it was not only capable to perform a full two-dimensional analysis of the properties in the channel area but also a significant speed-up and generally better numerical properties were obtained. E.g. for the calculation of a typical CMOS example, discretized at about 5000 nodes, the classical simulation took 55s CPU time per operation point and was reduced to 18s for the QM calculations. This simulations were performed on a Linux machine with a 650MHz Intel Celeron CPU. A possible reason for the speed up can be the smoother distributions when quantum confinement is taken into account.

2.4 Outlook

In this paper we presented a new macro-model for the simulation of the quantum confinement near the channel surface. A very good fit was obtained for the vertical carrier profile and together with the good numerical properties a deeper insight into the properties of future device generations is possible.

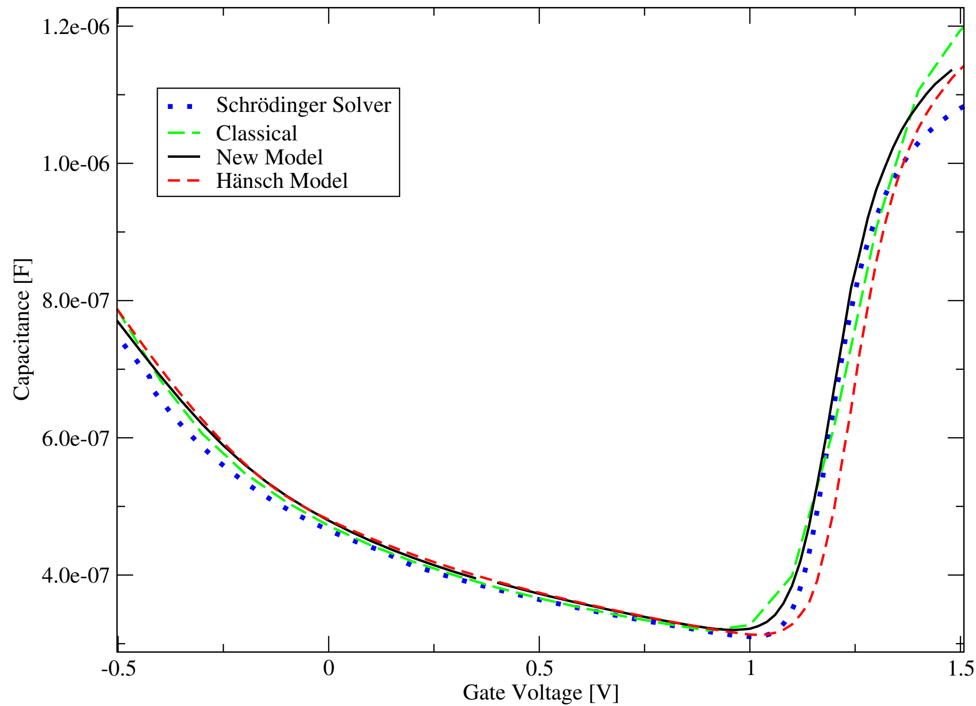

Figure 11: Comparison of C/V curves of a MOS capacitor obtained with different simulation approaches. The Hänsch model simulations were carried out with optimized offset z_0.

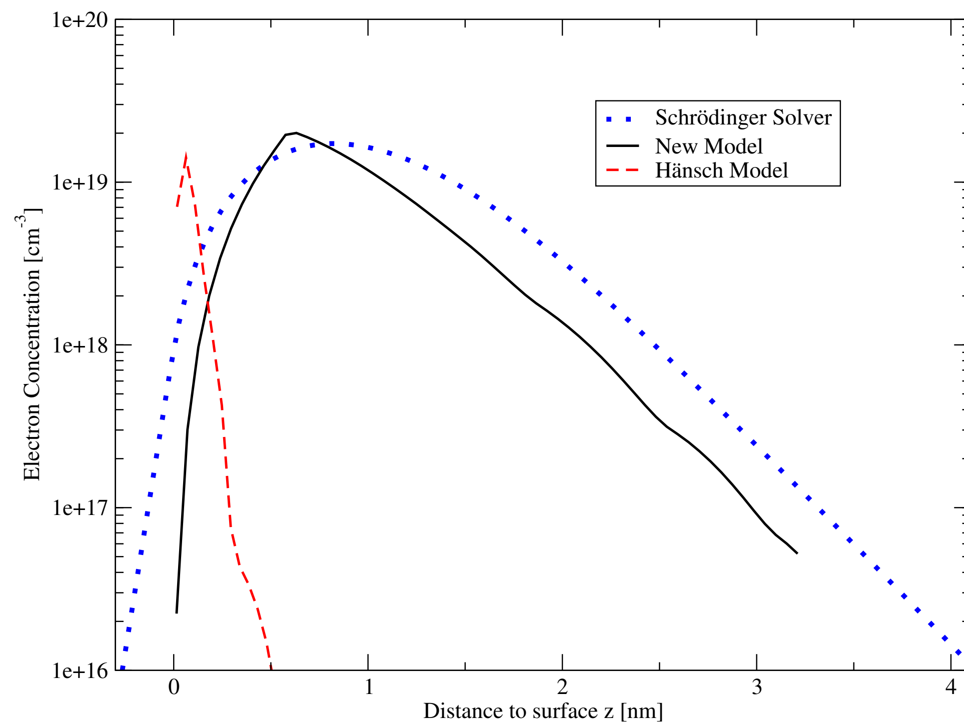

Figure 12: Comparison of carrier profiles. The operating points for the different models were chosen to obtain similar maximum concentrations. The same parameter sets as for the C/V calculations were applied.

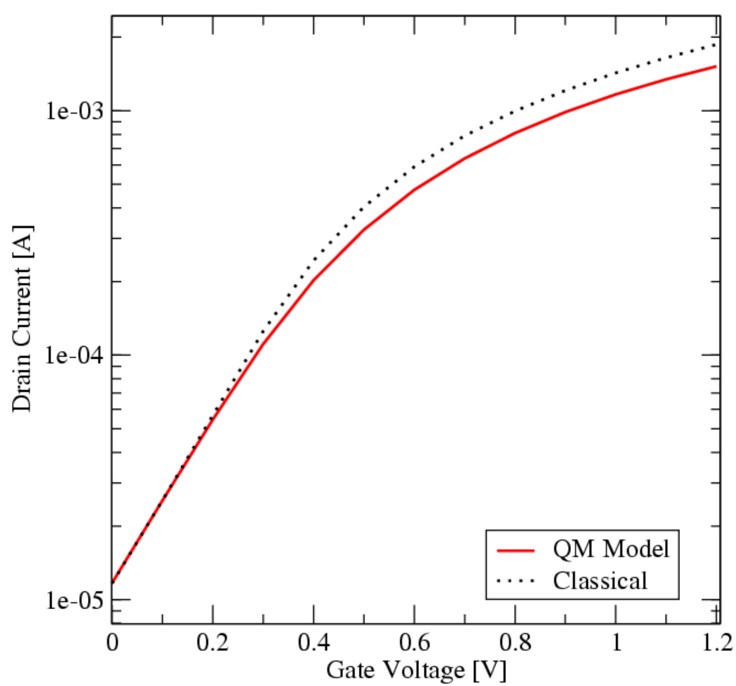

Figure 13: Transfer characteristics at of a 50nm NMOS at bias $U_{\mathrm{SD}} = 1\mathrm{V}$. The classical approach overestimates the drain current.

3 A Tunneling Model Based on a Non-Maxwellian Energy Distribution

3.1 Introduction

In state-of-the-art sub-quartermicron devices with gate oxide thicknesses around or below 2 nm the proper simulation of gate oxide tunneling currents is indispensable. Approaches to describe tunneling effects usually assume a Fermi-Dirac or cold Maxwellian shape of the electron energy distribution function (EED). This assumption is clearly violated in the channel of contemporary MOS devices where the deviation from the Maxwellian shape may reach several orders of magnitude due to a pronounced high energy tail. A correct modeling of the shape of the distribution function is thus necessary to reliably predict gate currents due to hot electron injection.

We report on a new formulation to describe hot electron injection through gate dielectrics. It is based on an expression which accounts for the non-Maxwellian shape of the electron energy distribution function. We use the first three even moments of the boltzmann equation n, T_n, and β_n found by the solution of a six moments transport model to describe the shape of the distribution function. Excellent agreement with results from rigorous Monte Carlo simulations and measurements is achieved.

3.2 Tunneling Model

Following[21] the hot electron gate current density can be derived from the Tsu-Esaki expression as

$$J_{\mathrm{g}} = q \cdot \int_0^\infty f(\mathcal{E}) \cdot g(\mathcal{E}) \cdot v_\perp(\mathcal{E}) \cdot T(\mathcal{E})\, d\mathcal{E} \tag{24}$$

where $f(\mathcal{E})$ is the electron energy distribution, $g(\mathcal{E})$ the density of states, $v_\perp(\mathcal{E})$ the electron velocity perpendicular to the interface, and $T(\mathcal{E})$ the tunneling probability. The integration is performed starting from the conduction band edge which serves as reference energy. The velocity perpendicular to the interface can be derived from the dispersion relation using the expression $v_\perp(\mathcal{E}) = 1/4 \cdot \partial\mathcal{E}/\partial p$ [21]. This approach offers the possibility to explicitly take the non-Maxwellian shape of the electron energy distribution function into account.

A simple model for the transmission coefficient can be derived using the WKB approximation for trapezoidal and triangular barriers:

$$T(\mathcal{E}) = \exp\left\{-4\frac{\sqrt{2m_{\mathrm{ox}}}}{3\hbar q F_{\mathrm{ox}}} \cdot \phi(\mathcal{E})\right\} \tag{25}$$

with F_{ox} being the electric field and m_{ox} the electron mass in the oxide. The function $\phi(\mathcal{E})$ is defined as

$$\phi(\mathcal{E}) = \begin{cases} (\Phi - \mathcal{E})^{3/2} & \Phi_0 < \mathcal{E} < \Phi \\ (\Phi - \mathcal{E})^{3/2} - (\Phi_0 - \mathcal{E})^{3/2} & \mathcal{E} < \Phi_0 \end{cases} \tag{26}$$

where Φ and Φ_0 are the upper and lower barrier heights, as shown in the inset in Fig. 14. The value of Φ_0 is calculated from $\Phi_0 = \Phi - q \cdot F_{\mathrm{ox}} \cdot t_{\mathrm{ox}}$ where q is the electron charge and t_{ox} denotes the gate oxide thickness. The transmission coefficient for different oxide thicknesses is depicted in Fig. 14 for silicon dioxide and several oxide thicknesses.

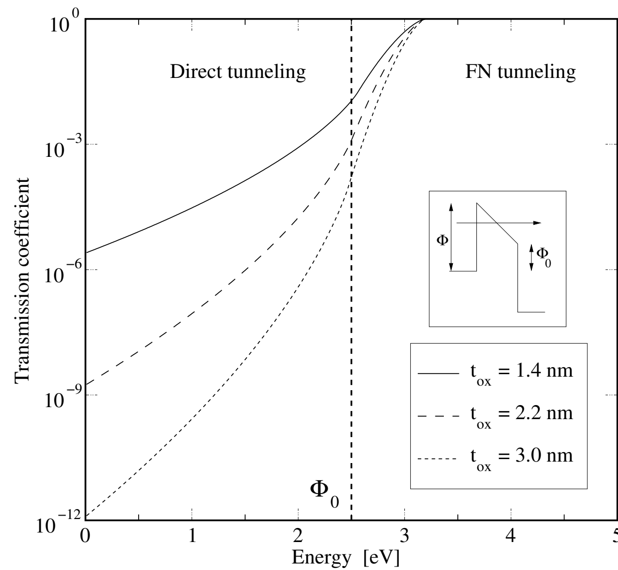

Figure 14: Transmission coefficient for different oxide thicknesses with $\Phi_0 = 2.5\,\text{eV}$.

3.3 Distribution Function

Various research deals with the problem of distribution function modeling for hot carriers in the channel region of a MOSFET [22, 23]. The problem arises from the fact that the assumption of a cold Maxwellian distribution function

$$f(\mathcal{E}) = A \cdot \exp\left(-\frac{\mathcal{E}}{k_\mathrm{B} \cdot T_\mathrm{L}}\right), \tag{27}$$

with T_L being the lattice temperature and A a normalization constant accounting for the Fermi energy, underestimates the high-energy tail of the electron energy distribution near the drain region. The straightforward approach is to use a heated Maxwellian distribution function

$$f(\mathcal{E}) = A \cdot \exp\left(-\frac{\mathcal{E}}{k_\mathrm{B} \cdot T_\mathrm{n}}\right) \tag{28}$$

where the lattice temperature T_L is simply replaced by the electron temperature T_n calculated from a suitable transport model.

We applied a Monte Carlo simulator employing analytical non-parabolic bands to check the validity of the heated Maxwellian approximation. The effect of electron-electron interaction, which increases the population of the high energy tail [22], was neglected in this study. Fig. 15 shows the contour lines of the heated Maxwellian EED in comparison to Monte Carlo results for a MOSFET device with a gate length of $L_\mathrm{g} = 180\,\mathrm{nm}$ at $V_\mathrm{DS} = V_\mathrm{GS} = 1\,V$. It can be clearly seen that the heated Maxwellian distribution (full lines) yields only poor agreement with the Monte Carlo results (dashed lines). The heated Maxwellian distribution overestimates the high-energy tail in the channel by several orders of magnitude. Furthermore, at the drain end of the channel hot electrons mix with cold electrons supplied from the drain region, which leads to an additional population of cold electrons which cannot be reproduced by this model.

A correct expression for the EED has to account for both effects, namely for the slope of the high energy tail along the channel and for the emerging population of cold electrons near the drain contact. Such an expression was proposed by Sonoda *et al.* [24], and an improved model has been suggested by Grasser *et*

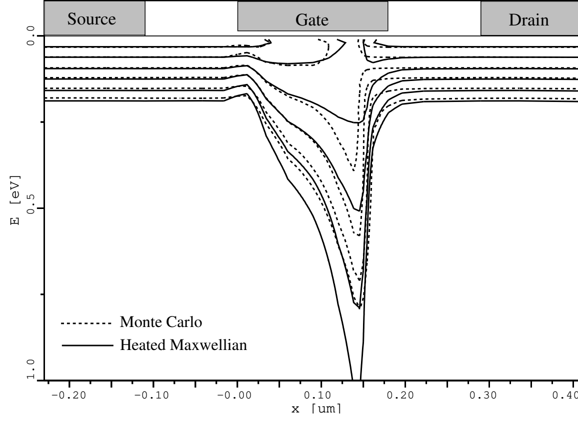

Figure 15: Heated Maxwellian EED compared to Monte Carlo results for a 180 nm MOSFET.

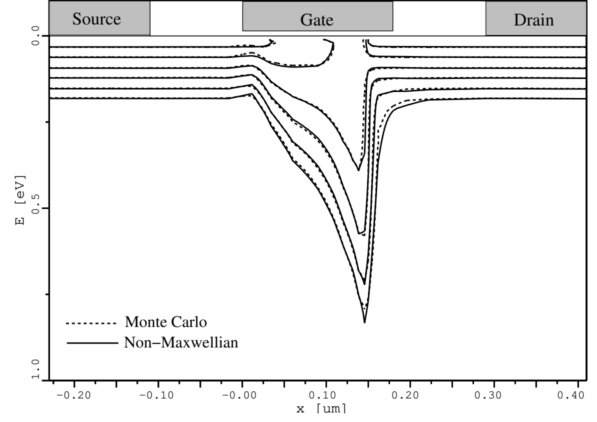

Figure 16: Non-Maxwellian EED compared to Monte Carlo results for a 180 nm MOSFET.

al.[25]:

$$f(\mathcal{E}) = A \left\{ \exp\left[-\left(\frac{\mathcal{E}}{a}\right)^b \right] + c \cdot \exp\left[-\frac{\mathcal{E}}{\mathrm{k_B} T_\mathrm{L}} \right] \right\}. \tag{29}$$

Here the pool of cold carriers in the drain region is correctly modeled by an additional cold Maxwellian subpopulation which leads to a reduced high-energy tail. The values of a, b, and c are derived from the solution variables of a six moments transport model using the procedure described in Ref. [25]. In Fig. 16 expression (29) is compared to Monte Carlo results showing excellent agreement all along the channel.

Note that only a six moments transport model can provide information about the kurtosis of the distribution function. The electron concentration, electron temperature and kurtosis are derived from

$$n = \langle 1 \rangle \tag{30}$$

$$\frac{3 k_\mathrm{B} T_\mathrm{n}}{2} = \langle \mathcal{E} \rangle \tag{31}$$

$$\frac{5 \beta_\mathrm{n}}{3} = \frac{\langle \mathcal{E}^2 \rangle}{\langle \mathcal{E} \rangle^2} \tag{32}$$

where the moments of the distribution function are defined as

$$\langle \Phi \rangle = \int_0^\infty \Phi \, f(\mathcal{E}) \, g(\mathcal{E}) \, d\mathcal{E}. \tag{33}$$

The value of the relative kurtosis in the bulk is used to locate the regions where the cold Maxwellian part of the distribution function emerges.

The value of the normalization constant A is calculated from the carrier concentration. This assures consistency of the model, in contrast to the normalization used by Hasnat *et al.* [26] which is independent of the carrier concentration and inevitably leads to erroneous results for both the carrier concentration and the electron temperature. Additionally, the population of cold carriers near the drain side of the channel can not be reproduced using Hasnat's model.

3.4 Results

For the evaluation of the tunneling model we apply our non-Maxwellian distribution function to the simulation of MOS transistors with varying gate lengths and oxide thicknesses. We simulated nMOS devices in on-state with $V_{\mathrm{GS}} = 1\,\mathrm{V}$ and $V_{\mathrm{DS}} = 1\,\mathrm{V}$. Gate lengths of 350 nm, 250 nm, 180 nm and 100 nm with gate oxide thicknesses of 3.4 nm, 3.0 nm, 2.6 nm, 2.2 nm, 1.8 nm, 1.4 nm, and 1.0 nm have been assumed. Gaussian source and drain doping peaks of $10^{20}\mathrm{cm}^{-3}$ with LDD extensions were used. In the following figures the results from Monte Carlo simulations will serve as reference.

Fig. 17 shows the integrand of expression (24) as a function of the electron energy for the case of a heated Maxwellian distribution and the non-Maxwellian distribution function (expressions (28) and (29)). The simulated device has a gate length of 100 nm and a gate oxide thickness of 3 nm. While at low energies the difference between the non-Maxwellian distribution function and the heated Maxwellian distribution seems to be negligible, the amount of overestimation of the incremental gate current density for the heated Maxwellian distribution reaches several orders of magnitude at 1 eV and peaks when the electron energy exceeds the barrier height. This spurious effect is clearly more pronounced for points at the drain end of the channel where the electron temperature is high.

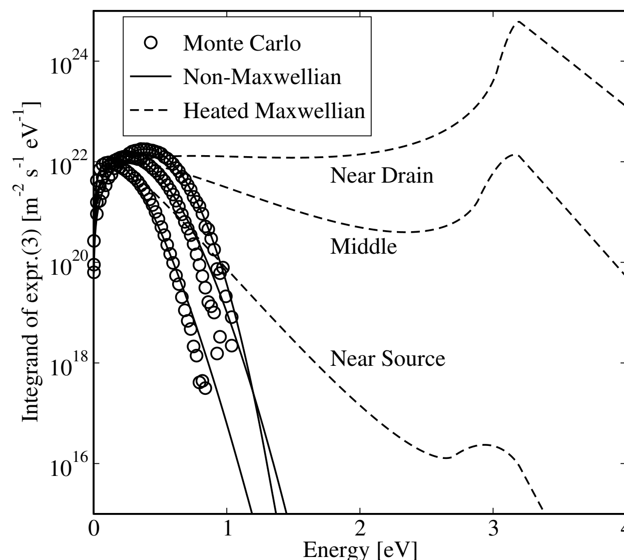

Figure 17: Integrand of (24) at different points in the channel of a 100 nm MOSFET.

The peak in the integrand for the heated Maxwellian EED results in an increased gate current density near the drain side of the gate contact. Fig. 18 shows the gate current density along the channel of a 180 nm gate length device for different oxide thicknesses. The heated Maxwellian EED heavily overestimates the gate current density near the drain contact, while the non-Maxwellian expression (29) correctly reproduces the Monte Carlo results.

In Fig. 19 the gate current density is depicted for the 100 nm and 180 nm device for an oxide thickness of 2.6 nm. While the non-Maxwellian distribution correctly reproduces the Monte Carlo results, the cold Maxwellian distribution leads to a sound underestimation reaching one order of magnitude, and the heated Maxwellian distribution predicts a much too high gate current density near the drain side of the channel. Due to the increasing electron temperature in the channel this effect is even more pronounced for smaller gate lengths.

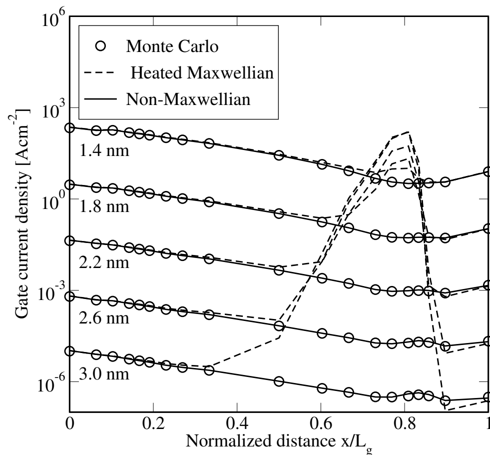

Figure 18: Gate current density along the channel of a 180 nm MOSFET.

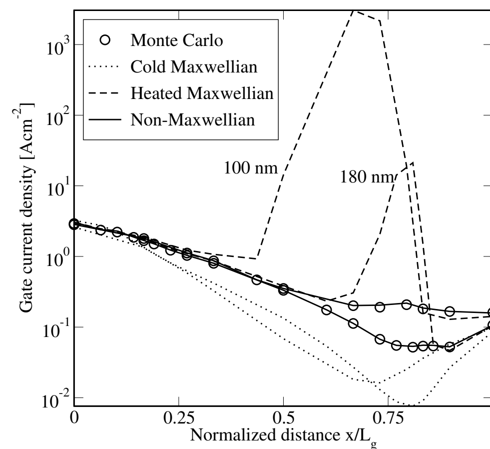

Figure 19: Gate current density along the channel of a 180 nm and a 100 nm MOSFET.

The simulation results have been compared to data reported in recent publications. For a MOSFET with zero drain-source voltage, the cold Maxwellian, heated Maxwellian, and non-Maxwellian model deliver of course the same results which are shown in Fig. 20. The measurement values were taken from Ref. [27]. The electron mass in the oxide was used as fitting parameter and an excellent fit for all oxide thicknesses was achieved with $m_{\mathrm{ox}} = 0.65\,m_0$. Note that the result is independent of the gate length since no drain-source bias was applied.

For the case of hot carriers, however, the heated Maxwellian distribution fails to reproduce even the qualitative behavior observed in measurements. Fig. 21 shows the gate current density as a function of the drain-source voltage for a 350 nm gate length, 1.8 nm oxide thickness MOSFET compared to measurement data presented in [28]. Perhaps due to differences in gate oxide thickness determination and the measurement setup, the gate oxide thickness had to be increased to give the same values for zero drain voltage as in Fig. 20. For increasing drain voltage, the electric field in the oxide is reduced which leads to a reduced transmission coefficient and lower gate current. The cold Maxwellian distribution again underestimates the amount of gate current, while The heated Maxwellian distribution overestimates the gate current density especially for high bias. While this effect is not so strong for the 350 nm device, it is clearly visible for a 180 nm device and, due to the increase of the electric field, it will even be higher for increased drain voltages. The non-Maxwellian model, however, correctly reproduces the measurements and shows reasonable results for the 180 nm device.

3.5 Conclusion

We presented a model to describe the hot-electron gate tunneling current by taking the non-Maxwellian shape of the electron energy distribution function into account. Conventional models which are based on the assumption of a cold or heated Maxwellian distribution fail to reproduce hot carrier tunneling in contemporary short-channel devices. The heated Maxwellian assumption delivers a much too high gate current density due to an overestimation of the high energy tail. Our model makes use of a recently developed non-Maxwellian expression for the distribution function which is based on a six-moments transport model. Using the new expression we can accurately reproduce Monte Carlo results and measurement data of a turned-on MOSFET.

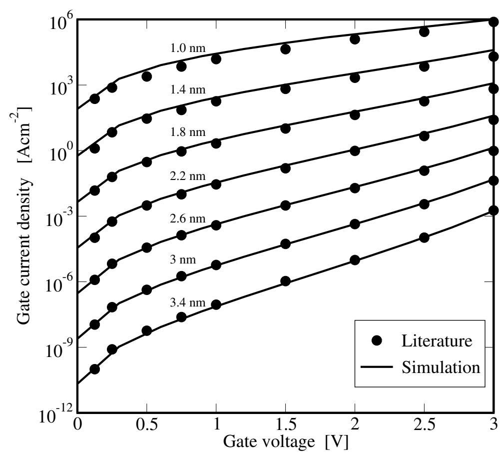

Figure 20: Gate current density for different MOS capacitors.

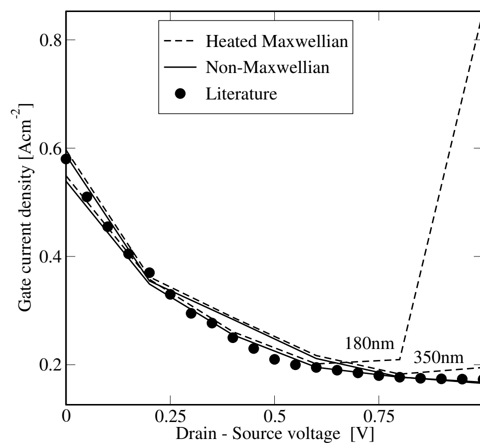

Figure 21: Gate current density for a 350 nm and a 180 nm device.

4 Investigation of the Electron Mobility in Strained $Si_{1-x}Ge_x$

4.1 Introduction

Materials which are compatible with the established Si technology are of particular interest if they allow to improve the transport properties of advanced devices. One such material is the alloy of Si and Ge parameterized by the Ge mole fraction *x*, which can be used both for $Si_{1-x}Ge_x$ active layers and $Si_{1-y}Ge_y$ substrates separately or together. If the Ge mole fraction in an active layer is different from that of the substrate, the resulting strain will cause changes in the band structure of the alloy and thus affects the kinetic properties of the material, in particular, the carrier mobility.

To obtain the proper electrical characteristics of SiGe devices it is necessary to have a correct mobility model including strain dependence. A rather good description of physics in semiconductors is possible by solving the Boltzmann transport equation using the Monte Carlo technique which allows semiclassical transport to be analyzed in a relatively complete form. It requires, however, accurate models for scattering rates which are in general functions of strain.

4.2 Strain Effects. Undoped Material

In this work we study the behavior of electron mobility in $Si_{1-x}Ge_x$ strained active layers grown on relaxed (001) $Si_{1-y}Ge_y$ substrate and the influence of scattering processes involving *L* valleys on electron mobility at high Ge mole fraction. The electron mobility at high Ge mole fraction strongly depends on the *L-X* scattering processes in strained SiGe alloys and may even have Si-like character in pure strained Ge, that is Ge with Si-like band structure.

Our analysis was performed by Monte Carlo simulations using the band structure reported in [29]. It consists of one conduction band and takes into account nonparabolicity and anisotropicity. Within this model the energy dependence on wave-vector is given by the following relation

$$E(\vec{k})(1+\alpha E(\vec{k})) = \frac{\hbar^2}{2}\left(\frac{k_1^2}{m_1}+\frac{k_2^2}{m_2}+\frac{k_3^2}{m_3}\right) \tag{34}$$

where α is a nonparabolicity parameter equal 0.5 for *X* valleys in Si and 0.3 for *L* valleys in Ge $m_1 = m_2 = m_t$ are transverse components of the effective mass tensor and $m_3 = m_l$ is the longitudinal component. As we investigate the low-field electron mobility, the analytical description will be sufficient to describe the main characteristic features of strained SiGe active layers. To study high-field transport phenomena, one has to use the full-band representation of the band structure.

The model of Rieger and Vogl [30] was adopted for the effective masses in strained SiGe. This model gives the effective masses versus Ge mole composition in the active layer and the substrate:

$$m^*(x,y) = \begin{pmatrix} 1, & (x-y)\,, & (x-y)^2 \end{pmatrix} \mathbf{W} \begin{pmatrix} 1 \\ (x+y) \end{pmatrix}$$

where $\mathbf{W}$ contains parameterized transverse and longitudinal effective masses for the perpendicular and parallel *X*-valleys, and *x* and *y* denote the Ge mole fractions of the active layer and the substrate, respectively.

The linear deformation potential theory [31] was used to calculate the splittings of the valleys. The shift of the *i*-th valley is expressed as

$$\Delta E_c^i = \Xi_d \cdot \mathbf{Tr}(\epsilon) + \Xi_u \cdot \mathbf{a}_i^T \cdot \epsilon \cdot \mathbf{a}_i \tag{35}$$

where ϵ is the strain tensor, $\mathbf{a}_i$ is a unit vector parallel to the $\vec{k}$ vector of the *i*-th valley, Ξ_d and Ξ_u are the dilatation and shear deformation potentials. The deformation potential tensor which is diagonal along the Δ axis, has two independent components Ξ_l^{Δ}, Ξ_t^{Δ} and $\Xi_u = \Xi_l - \Xi_t$, $\Xi_d = \Xi_t$. For the shift of the mean energy of the conduction band minima one obtains from ((35)):

$$\Delta E_c = (\Xi_d + \frac{1}{3}\Xi_u) \cdot \mathbf{Tr}(\epsilon). \tag{36}$$

For the uniaxial strain along [001] the strain tensor has diagonal form in the principle coordinate system:

$$\epsilon = \begin{bmatrix} \epsilon_{xx} & 0 & 0 \\ 0 & \epsilon_{yy} & 0 \\ 0 & 0 & \epsilon_{zz} \end{bmatrix} \tag{37}$$

where $\epsilon_{xx} = \epsilon_{yy} = \epsilon_{||}$ and $\epsilon_{zz} = \epsilon_{\perp}$ are expressed in terms of the lattice constants of strained active layer $a_{||}$, $a_{\perp}$ and relaxed material a_0 by $\epsilon_{||} = (a_{||} - a_0)/a_0$, $\epsilon_{\perp} = (a_{\perp} - a_0)/a_0$. The energy splitting within a given valley is the difference of two components: the shift of the mean energy of the band minima and the shift of an individual direction minimum for this valley. The first component is given by expression ((36)), the second one can be obtained from ((35)). From these relations it follows:

$$\Delta E_c^{[001]} = \Delta E_c^{[00\bar{1}]} = \frac{2}{3} \cdot \Xi_u^{\Delta} \cdot (\epsilon_{zz} - \epsilon_{xx}) \,, \tag{38}$$

$$\Delta E_c^{[100]} = \Delta E_c^{[\bar{1}00]} = \Delta E_c^{[010]} = \Delta E_c^{[0\bar{1}0]} = -\frac{1}{3} \cdot \Xi_u^{\Delta} \cdot (\epsilon_{zz} - \epsilon_{xx}) \tag{39}$$

This means that under the uniaxial strain along [001] the minima of the conduction band at Δ near the *X*-points are shifted with respect to the mean energy. Two valleys along the strain direction are shifted in the opposite direction to the shift of the in-plane valleys. It is easy to see from the derivation given above that the *L*-valleys remain degenerate under [001] strain. However, these valleys are split under [111] strain making the Monte Carlo analysis more complex due to the necessity to modify all scattering processes which involve the *L*-valleys.

As the valleys on the Δ axes are not equivalent any longer, all scattering mechanisms which include the *X*-valleys as an initial or final valley have to be changed accordingly. This point is important for the *L-X* and *f-type X-X* scattering processes.

Since we study the low field electron mobility in undoped strained bulk SiGe, a Monte Carlo algorithm for zero field which was developed for the nondegenerate case [32] is applied. This algorithm is very efficient and allows to obtain the whole mobility tensor. For our simulation we consider phonon scattering of both types intravalley and intervalley. Alloy scattering is taken into account as proposed in [33].

4.3 Inclusion of Ionized Impurity Scattering Rate

We adopted the model of ionized impurity scattering which was developed for majority electrons in Si [34]. This model takes into account several effects such as momentum dependent screening, multi-potential scattering, the second Born correction, and the Pauli exclusion principle. Plasmon scattering is included as an additional mechanism. A method which reduces the number of small-angle scattering events [35] is applied. It gives the same momentum relaxation time and represents an isotropic process.

In order to account for minority carrier transport the same model is used, but without momentum-dependent screening, Pauli exclusion principle and plasmon scattering. In addition, screening by holes

is considered and modeling of the valence bands is taken into account as suggested in [36, 37]. The presence of strain requires the proper calculation of the Fermi energy in strained SiGe active layers. As opposed to the undoped material, the Fermi energy in the strained layer is now defined by the following nonlinear equation:

$$n = N_{c_\perp} \cdot \mathcal{F}_{1/2}\left(\frac{E_f - E_c - \Delta E_\perp}{k_B T}\right) + N_{c_\parallel} \cdot \mathcal{F}_{1/2}\left(\frac{E_f - E_c - \Delta E_\parallel}{k_B T}\right) \quad (40)$$

where n is the electron concentration, $N_{c_\perp}$, $N_{c_\parallel}$ are the effective density of states for the split set of valleys of expressions ((38)) and ((39)) respectively, $\Delta E_\perp$, $\Delta E_\parallel$ are the energy shifts for the same set of valleys, $\mathcal{F}_{1/2}$ is the Fermi integral of respective order, $\frac{E_f - E_c}{k_B T}$ is the reduced Fermi energy which has to be found. This equation is solved by a Newton nonlinear iteration, and Maxwell-Boltzmann statistics is used to define the initial guess.

Also the expression for the inverse screening length changes in the presence of strain,

$$\beta_s^2 = \frac{e^2}{\varepsilon_s \varepsilon_0 k_B T}\left(N_{c_\perp}\mathcal{F}_{-1/2}(\eta_\perp) + N_{c_\parallel}\mathcal{F}_{-1/2}(\eta_\parallel)\right) \quad (41)$$

where ε_s is the relative dielectric permittivity of the semiconductor and $\eta_i = \frac{E_f - E_c - \Delta E_i}{k_B T}$.

Additionally, if the case of momentum-dependent screening is considered, the screening function has to be modified. The proper modification for this effect follows from the formula for the dielectric function which under the strain considered here takes the following form:y

$$\varepsilon(q) = \varepsilon(0)\left(1 + \frac{1}{q^2}(\beta_{s_\perp}^2 G(\xi, \eta_\perp) + \beta_{s_\parallel}^2 G(\xi, \eta_\parallel))\right) \quad (42)$$

where q is the momentum transfer, G is the screening function, $\xi = \frac{\hbar^2 q^2}{8m^* k_B T}$ and $\beta_{s_i}^2 = \frac{e^2}{\varepsilon_s \varepsilon_0 k_B T} N_{c_i} \cdot \mathcal{F}_{-1/2}(\eta_i)$.

4.4 Results

In Fig. 22 the behavior of electron mobility in an unstrained SiGe layer and a strained SiGe layer on Si is shown. The experimental data are taken from [38]. As can be seen from this figure, mobility in the strained case has Si like character over the whole range of Ge mole fraction. It is related to the increase of the biaxial compressive strain which at high x makes the four in-plane valleys move down getting equal or even lower than the L-valleys. Fig. 24 shows the mobility for strained SiGe on relaxed SiGe substrate. In this case at the beginning of the curves the four parallel valleys are shifted up. As the Ge mole fraction in the layer increases, the tensile strain decreases and turns at the point $x = y$ (unstrained point) into compressive strain. In this case the four in-plane X-valleys move down but they are always higher than the L-valleys in such conditions. The results for the corresponding in-plane mobility are presented in Fig. 25. The doping dependence of the perpendicular and the in-plane components is shown in Fig. 26 and Fig. 27. The increase of the perpendicular component at high doping concentrations is explained by the dominance of ionized impurity scattering over alloy scattering. This is also clearly seen in Fig. 28 while the in-plane component Fig. 29 does not have this increase as it follows from the energy splitting.

4.5 Conclusion

We presented the Monte Carlo simulations of strained SiGe active layers grown on relaxed (001) SiGe substrate. The physical model based on the analytical representation of the band structure includes strain

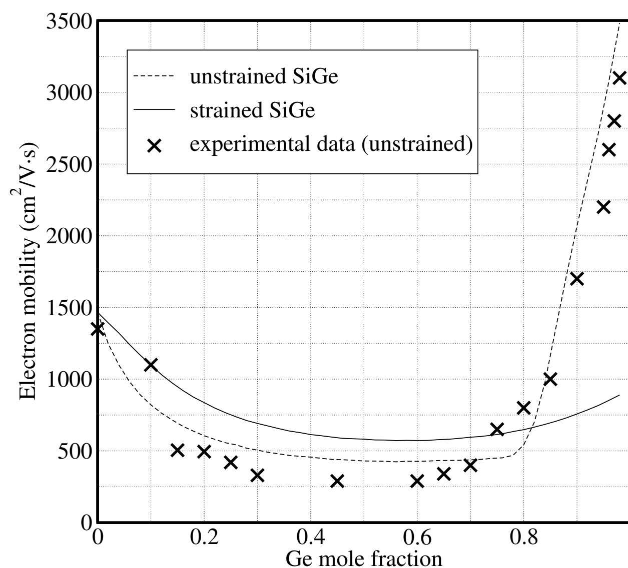

Figure 22: Electron mobility perpendicular to the interface in $Si_{1-x}Ge_x$ on a Si substrate at 300K.

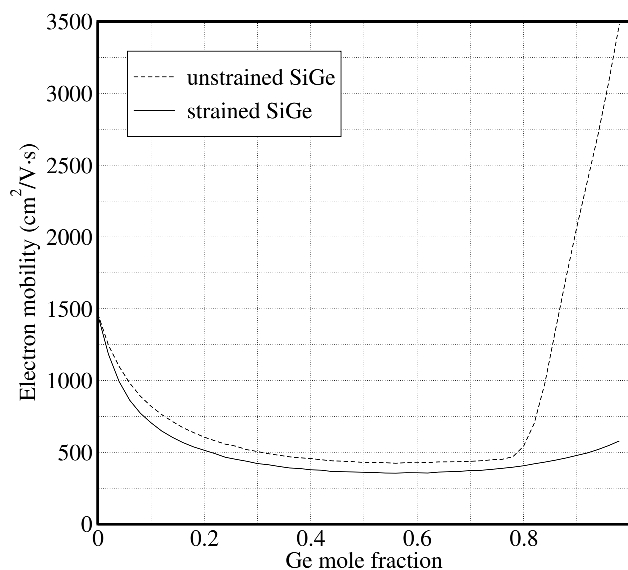

Figure 23: Electron mobility parallel to the interface in $Si_{1-x}Ge_x$ on a Si substrate at 300K.

effects for undoped and doped cases and allows important effects in highly doped semiconductors to be taken into account. The low-field electron mobility of strained SiGe layers including high Ge mole fraction was calculated. A good agreement of the obtained results with the experiment is achieved.

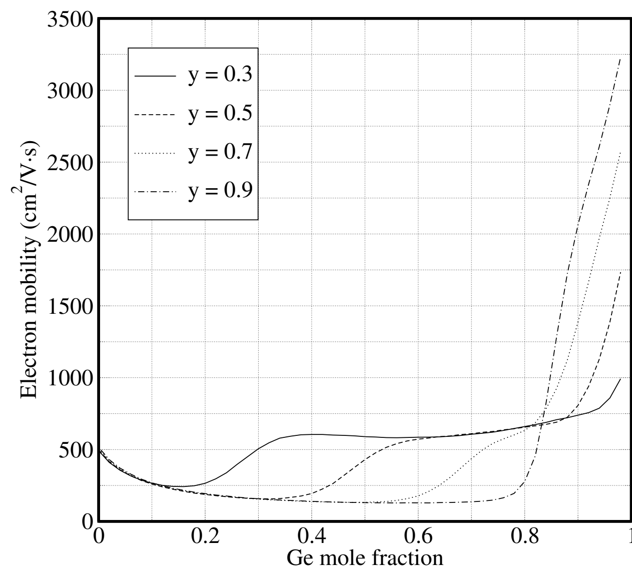

Figure 24: Electron mobility perpendicular to the interface in $Si_{1-x}Ge_x$ on a $Si_{1-y}Ge_y$ substrate at 300K.

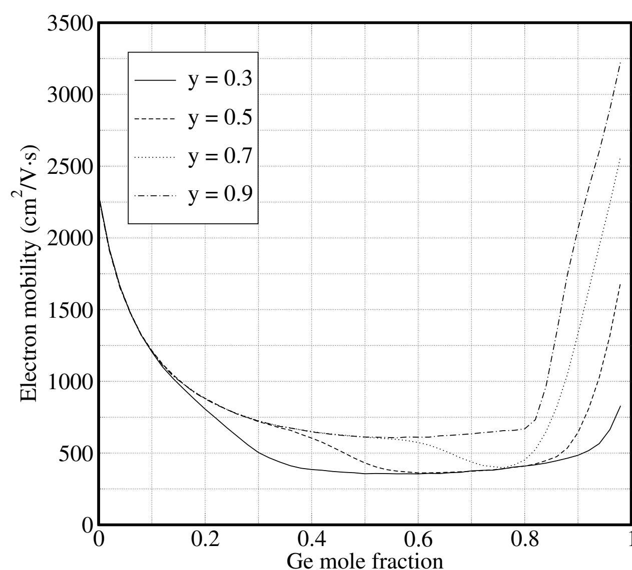

Figure 25: Electron mobility parallel to the interface in $Si_{1-x}Ge_x$ on a $Si_{1-y}Ge_y$ substrate at 300K.

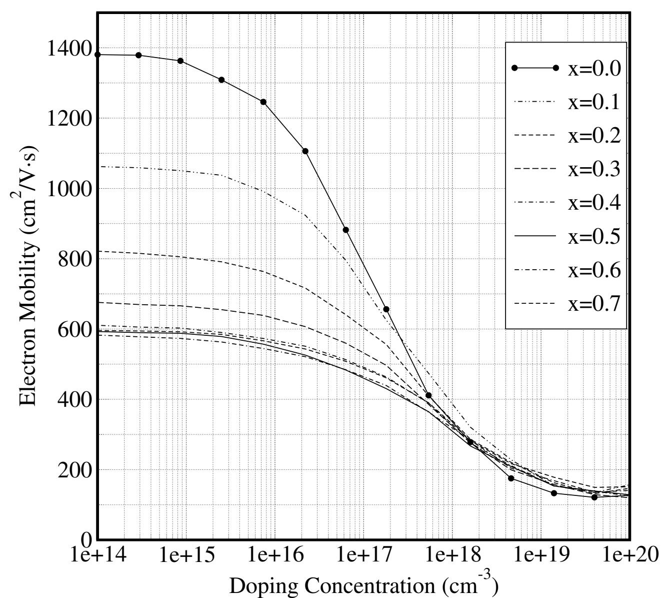

Figure 26: Doping dependence of electron mobility perpendicular to the interface in $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ on a Si substrate at 300K.

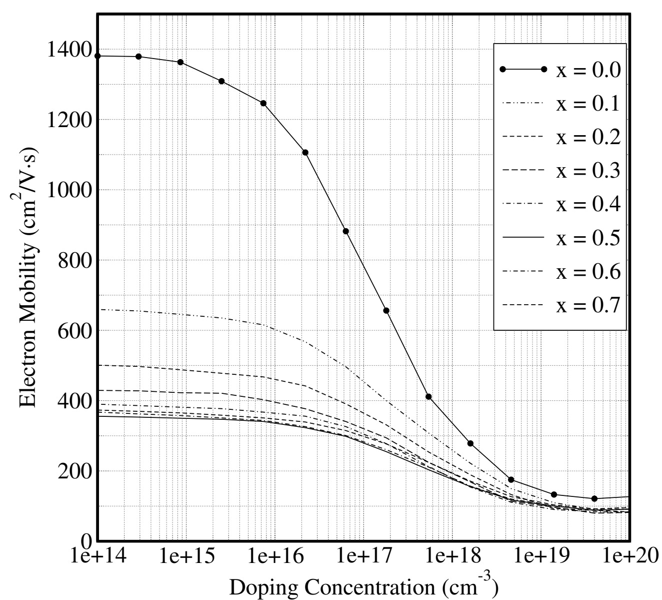

Figure 27: Doping dependence of electron mobility parallel to the interface in $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ on a Si substrate at 300K.

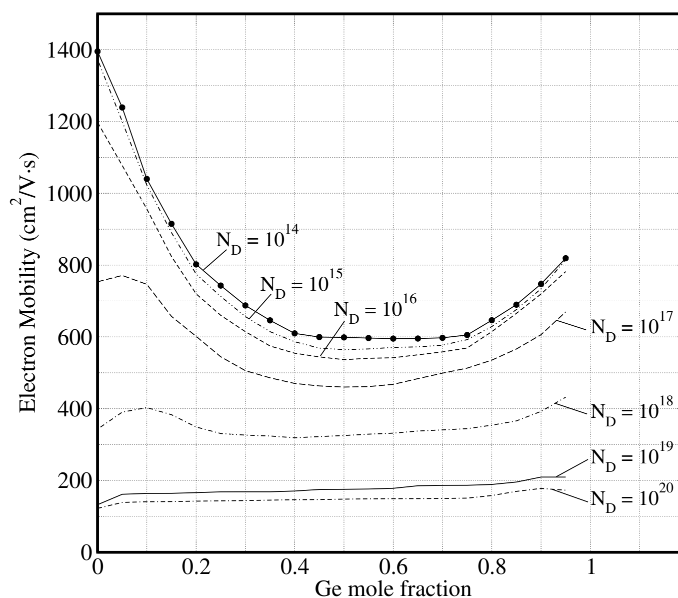

Figure 28: Ge composition dependence of electron mobility perpendicular to the interface in $Si_{1-x}Ge_x$ on a Si substrate for different doping levels at 300K.

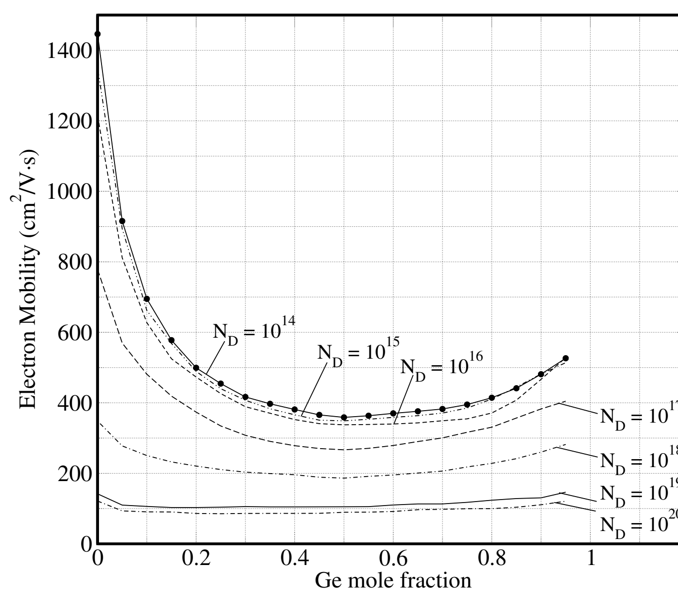

Figure 29: Ge composition dependence of electron mobility prallel to the interface in $Si_{1-x}Ge_x$ on a Si substrate for different doping levels at 300K.

5 Effects of Stress-Induced Bandgap Narrowing

The effects of bandgap narrowing due to stress generated during Shallow Trench Isolation (STI) are analyzed. Reverse-bias junction leakage and capacitance measurements are correlated to results from device simulation. A locally varying stress dependent bandgap model is implemented to understand the influence of stress effects. Increases in both junction capacitance and leakage agree well with experiments. Results of leakage and capacitance on the 0.18 μm and 0.09 μm technologies indicate that effects of process induced stress on device behavior need careful attention.

5.1 Introduction

Process induced stress occurs at several steps in the manufacturing of integrated circuits, for example, due to sharp trench corners at the top and bottom of the shallow trench, which is enhanced during subsequent liner oxidation, or due to the deposition of spacer nitride or inter-metal dielectric. In many cases, stress results in undesirable negative effects on device performance. In particular, the STI integration method has a strong effect on reverse-bias junction leakage [39]-[40]. Studies by Smeys et al. [41] indicated that, even in the absence of etch induced traps in the depletion region, oxidation stress caused junction leakage. They attributed this increased generation current to stress-induced bandgap narrowing. This leakage current increased with decreasing active area pitch and their model predicted an exponential dependence of leakage current on stress. Studies done by Gopinath et al. [42] on different liner oxidation schemes indicated a strong correlation between reverse-bias junction leakage and stress due to liner oxidation. Recently, the impact of stress-induced bandgap narrowing on junction capacitance was reported by [43]. It was shown that, in addition to leakage, junction capacitance also increases due to its dependence on bandgap energy. However, none of the studies reviewed so far quantitatively simulated the effects of bandgap narrowing on device characteristics.

This is largely due to the fact, that bandgap energy in the commonly used device simulators is treated by a global variable while the stress is a local grid based variable. Relevant information about stress and its components can be obtained as a result from process simulation, e.g. with TSUPREM4 [44]. Suitable modifications to use this important information have been made to MINIMOS-NT [45] in order to, consequently, correlate process simulation, device simulation, and experiments. Stress simulated by TSUPREM4 is used as input to MINIMOS-NT, which maps these changes into local variations of the bandgap. The device simulations reflect these variations and the correlation of the results with experiments are quite satisfactory.

5.2 Experiments

Devices from the 180 nm and 90 nm technology node were used to compare the simulations with measurements. The former experiment involved a combination of liner oxidation (15 nm and 30 nm) and liner undercut targets (17.5 nm, 90 nm, and no undercut) as listed in Table 1 and further detailed in [5].

The 90 nm node experiments were done to understand the effect of trench bottom stress on device behavior. This study was performed for trench depths of 0.17 μm, 0.2 μm and 0.23 μm and the reverse-bias junction capacitance and leakage were analyzed. This study also involved the use of different background well species (Boron and Indium for p-well, and Arsenic and Phosphorus for n-well) to gauge the effect of background doping on well junction behavior [43]. Both studies indicated a strong correlation between stress as reported by TSUPREM and reverse-bias junction characteristics. In order to enable two-dimensional study of the stress induced changes, long perimeter intensive diodes (200 fingers of 0.99 μm width and

Split No.	Liner oxide target [nm]	Nitride undercut target [nm]
1	15	0
2	15	9
3	15	17.5
4	30	0
5	30	9
6	30	17.5

Table 1: Liner and undercut splits

500 μm length) were used for correlating simulations with experiments. Simulating a section of the long diodes in the width direction reduces the three-dimensional problem to two dimensions. In the case of the 90 nm node, corner intensive devices, 27000 rectangles of 1.22 μm $\times$ 1.52 μm resulting in a total area of $5 \times 10^4 \mu\text{m}^2$ were used and simulated along the shorter dimension.

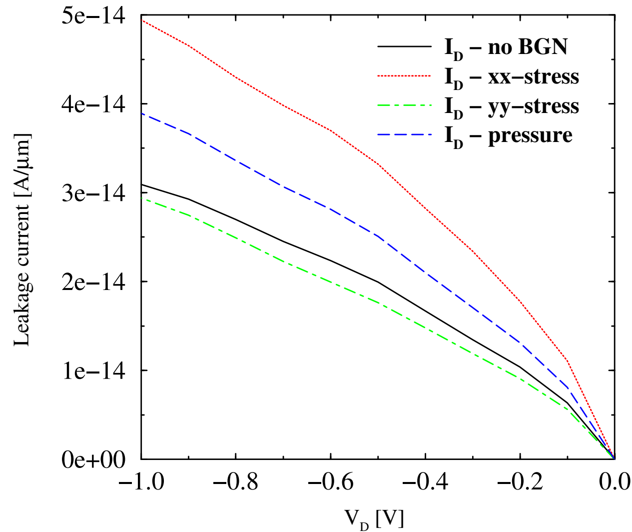

Figure 30: Reverse-bias junction characteristics of a p^+/n junction diode with varying stress.

5.3 Results and Discussion

The bandgap narrowing equation shown in (6) was used to implement a locally varying bandgap in MINIMOS-NT. The simulations of a simple p^+/n junction diode with and without the model are shown in Fig. 1. It can be seen that the reverse-bias junction leakage increases with compressive stress. An understanding of the stresses generated from STI integration can be seen in Fig. 2 for device with trench depth of 0.2 μm. The S_{xx} component of the stress shows very high compressive values at the bottom of the trench and high tensile values at the top. The effect of bandgap reduction on the generation rate can be seen in Fig. 3. It can be clearly seen that the generation current in the depletion region has the highest magnitude along the high-stress STI sidewall areas.

Fig. 4 shows the comparison of experimental (error bars represent 95% confidence intervals [46]) and simulated values of reverse-bias p^+/n junction leakage. Stress predicted by TSUPREM4 simulations for the experimental conditions was used to model the bandgap in MINIMOS-NT. It can be seen that the

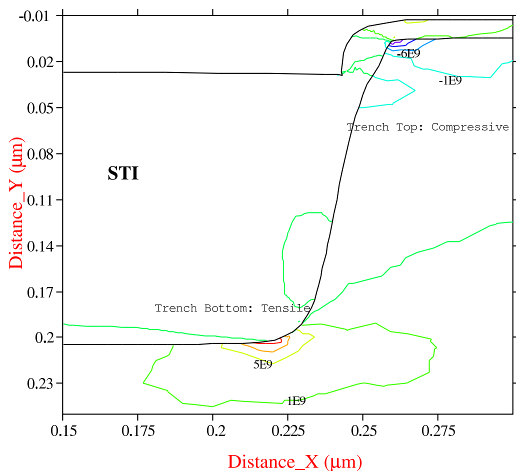

Figure 31: TSUPREM4 simulation of S_{xx} component of stress for trench depth of 0.2 μm. Note the high stress regions in the bottom and top of the trench.

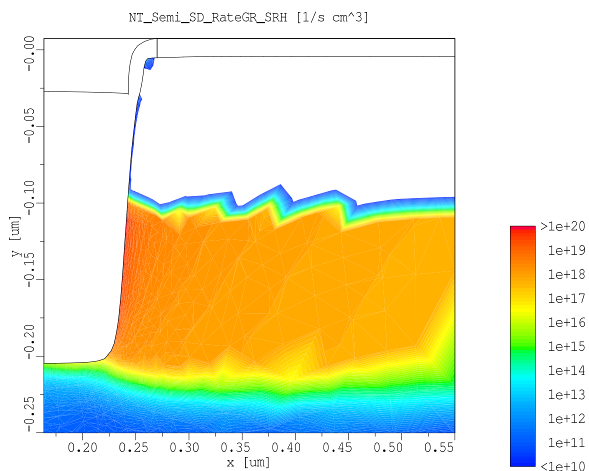

Figure 32: Generation rate [s^{-1} cm^{-3}]: Note the higher surface generation at the high-stress trench region.

bandgap narrowing model accurately reproduces both the magnitude and trend of the experimental data. However, the rate of change of leakage for the different process conditions is somewhat stronger than the predicted by device simulation. Fig. 5 shows the same result for n^+/p perimeter diodes for the six different splits. Again, the bandgap reduction simulations show good agreement with the median values of the measurement.

Capacitance simulations using the bandgap narrowing model were performed for 90 nm junctions with varying trench depths. The experimental and simulated results agree well as shown in Fig. 6. The simulated stress was higher for the shallower trench depth case [43] which in turn leads to larger capacitance and higher junction leakage. The bandgap narrowing model in MINIMOS-NT accurately reflects this behavior. The reverse-bias junction leakage for the same structure is shown in Fig 7. While the simulator models the trend correctly, the actual values are higher than measured. However, the studies confirm that, in cases where feasible (due to limitations of gap fill), a deeper trench is preferable since the depletion region will span less of the high stress trench bottom and will lead to improved junction behavior.

It should be noted here that only stresses generated during STI liner oxidation were simulated and transferred to MINIMOS-NT for device simulations. However, the actual device undergoes many thermal cycles which lead to further stresses (and relief) during processing. The version of TSUPREM used did not carry over the stress components for the entire process sequence, hence it was decided to conduct this study with only the STI part of the stress since all the process splits were at the STI module.

5.4 Conclusions

A locally varying bandgap model is introduced into the device simulator MINIMOS-NT. The bandgap is then varied according to process-induced stresses and the reverse-bias junction characteristics are simulated. The results correlate well with devices fabricated on the 180 nm and 90 nm nodes.

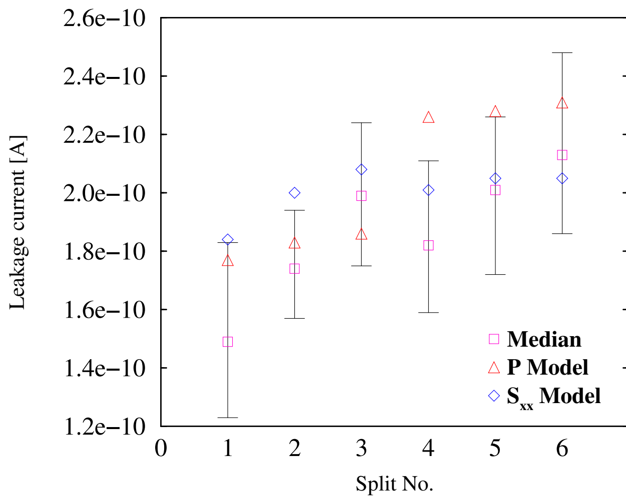

Figure 33: Reverse-bias junction leakage [A]: Results for p^+/n perimeter intensive diodes at -3.3 V bias. The error bars represent 95% confidence levels.

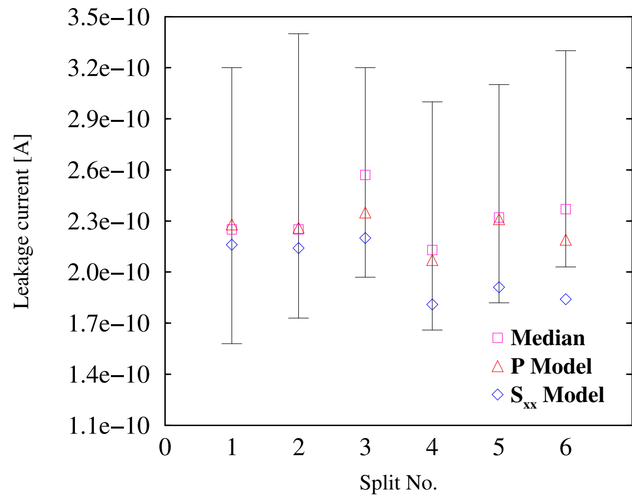

Figure 34: Reverse-bias junction leakage [A]: Results for n^+/p perimeter intensive diodes at 3.3 V bias. Measurements have higher scatter than their p^+/n counterparts.

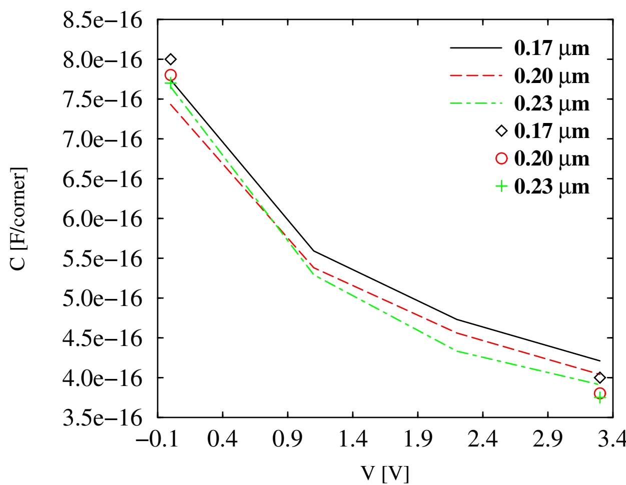

Figure 35: Reverse-bias junction capacitance [F/corner]: Simulations (lines) and measurements (symbols) for varying trench depths for n^+/p corner intensive diodes.

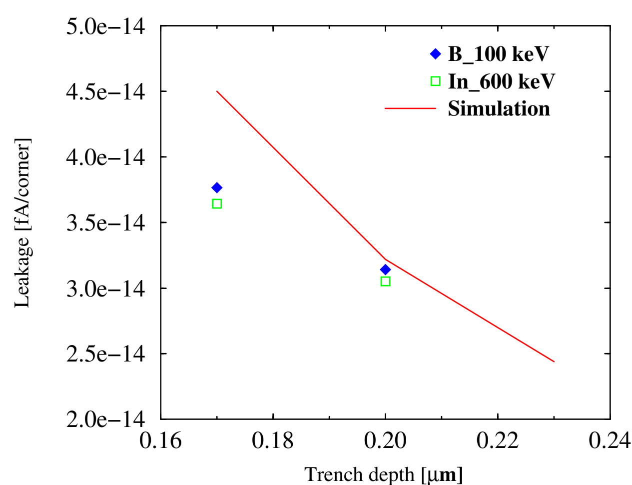

Figure 36: Reverse-bias junction leakage [A/corner]: Simulation (line) and measurements (symbols) for varying trench depths for n^+/p corner intensive diodes.

References

[1] D. R. Fridline and A. F. Bower. Influence of anisotropic surface diffusivity on electromigration induced void migration and evolution. *J. Appl. Phys.*, 85(6):3168–3174, 1999.

[2] M. Mahadevan and R. Bradley. Simulations and theory of electromigration-induced slit formation in unpassivated single-crystal metal lines. *Physical Review B*, 59(16):11037–11046, 1999.

[3] D. N. Bhate, A. Kummar, and A. F. Bower. Diffuse interface model for electromigration and stress voiding. *J. Appl. Phys.*, 87(4):1712–1721, 2000.

[4] M. Mahadevan, R. Bradley, and J. M. Debierre. Simulation of an electromigration-induced edge instability in single-crystal metal lines. *Europhys. Lett.*, 45:680–685, 1999.

[5] Z. Suo, W. Wang, and M. Yang. Electromigration instability: Transgranular slits in interconnects. *Appl. Phys. Lett*, 64(15):1944–1946, 1994.

[6] Z. Suo and W. Wang. Diffusive void bifurcation in stressed solid. *J. Appl. Phys.*, 76(6):3410–3421, 1994.

[7] Paul S. Ho. Motion of Inclusion by a Direct Current and a Temperature Gradient. *J. Appl. Phys.*, 41(1):64–68, 1970.

[8] M. Mahadevan and R. Bradley. Phase field model of surface electromigration in single crystal metal thin films. *Physica D*, 126:201–213, 1999.

[9] J. Blowey and C. Elliott. The Chan-Hilliard gradient theory for phase separation with non-smooth free energy, Part II: Numerical analysis. *European Journal of Applied Mathematics*, 3:147–176, 1992.

[10] J. Blowey and C. Elliott. The Chan-Hilliard gradient theory for phase separation with non-smooth free energy, Part I: Mathematical analysis. *European Journal of Applied Mathematics*, 2:233–279, 1991.

[11] C. Elliot and J. Ockendon. *Weak and Variational Methods for Moving Boundary Problems*. Pitman Publishing Inc., 1981.

[12] R. Sabelka. *"Dreidimensionale Finite Elemente Simulation von Verdrahtungsstrukturen auf Integrierten Schaltungen"*. Dissertation, Technische Universität Wien, 2001.

[13] I. Kossacky. A recursive approach to local mesh refinement in two and three dimensions. *J. Comput. Appl. Math.*, 55:275–288, 1994.

[14] E. Arzt, O. Kraft, W. D. Nix, and J. E. Sanchez. Electromigration failure by shape change of voids in bamboo lines. *J. Appl. Phys.*, 76(3):1563–1571, 1994.

[15] W. Hänsch, T. Vogelsang, R. Kircher, and M. Orlowski. Carrier Transport Near the Si/SiO_2 Interface of a MOSFET. *Solid-State Electron.*, 32(10):839–849, 1989.

[16] Z. Yu, R. Dutton, and R. Kiehl. Circuit/Device Modeling at the Quantum Level. In *IWCE-6*, pp 222–229, Osaka, Japan, 1998.

[17] M. Van Dort, P. Woerlee, and A. Walker. A Simple Model for Quantisation Effects in Heavily-Doped Silicon MOSFETs at Inversion Conditions. *Solid-State Electron.*, 37(3):411–414, 1994.

[18] P. Chow. Computer Solutions to the Schrödinger Equation. *Am.J.Phys.*, 40:730–734, 1972.

[19] T. Binder, K. Dragosits, T. Grasser, R. Klima, M. Knaipp, H. Kosina, R. Mlekus, V. Palankovski, M. Rottinger, G. Schrom, M. Stockinger, and S. Selberherr. *MINIMOS-NT User's Guide*. Institut für Mikroelektronik, Vienna, 1998.

[20] C. A. Richter, E. M. Vogel, A. M. Hodeg, and A. R. Hefner. Differences Between Quantum-Mechanical Capacitance-Voltage Simulators. In *Simulation of Semiconductor Processes and Devices*, Athens, Greece, 2001.

[21] C. Fiegna, F. Venturi, M. Melanotte, E. Sangiorgi, and B. Riccó. Simple and Efficient Modeling of EPROM Writing. *IEEE Trans.Electron Devices*, 38(3):603–610, 1991.

[22] A. Abramo and C. Fiegna. Electron Energy Distributions in Silicon Structures at Low Applied Voltages and High Electric Fields. *J.Appl.Phys.*, 80(2):889–893, 1996.

[23] D. Cassi and B. Riccó. An Analytical Model of the Energy Distribution of Hot Electrons. *IEEE Trans.Electron Devices*, 37(6):1514–1521, 1990.

[24] K.-I. Sonoda, M. Yamaji, K. Taniguchi, C. Hamaguchi, and S. T. Dunham. Moment Expansion Approach to Calculate Impact Ionization Rate in Submicron Silicon Devices. *J.Appl.Phys.*, 80(9):5444–5448, 1996.

[25] T. Grasser, H. Kosina, C. Heitzinger, and S. Selberherr. Characterization of the Hot Electron Distribution Function Using Six Moments. *J.Appl.Phys.*, 91(6):3869–3879, 2002.

[26] K. Hasnat, C.-F. Yeap, S. Jallepalli, S. A. Hareland, W.-K. Shih, V. M. Agostinelli, A. F. Tasch, and C. M. Maziar. Thermionic Emission Model of Electron Gate Current in Submicron NMOSFETs. *IEEE Trans.Electron Devices*, 44(1):129–138, 1997.

[27] A. Shanware, J. P. Shiely, and H. Z. Massoud. Extraction of the Gate Oxide Thickness of N- and P-Channel MOSFETs Below 20 Å from the Substrate Current Resulting from Valence-Band Electron Tunneling. In *Proc.IEDM Tech.Dig*, pp 815–818, 1999.

[28] S. Schwantes and W. Krautschneider. Relevance of Gate Current for the Functionality of Deep Submicron CMOS Circuits. In *European Solid-State Device Research Conference*, pp 471–474, Nuremberg, Germany, 2001.

[29] C. Jacoboni and L. Reggiani. The Monte Carlo Method for the Solution of Charge Transport in Semiconductors with Applications to Covalent Materials. *Rev. Mod. Phys.*, 55(3):645–705, 1983.

[30] M. M. Rieger and P. Vogl. Electronic-Band Parameters in Strained $Si_{1-x}Ge_x$ Alloys on $Si_{1-y}Ge_y$ Substrates. *Physical Review B*, 48(19):14276–14287, 1993.

[31] C. Herring and E. Vogt. Transport and Deformation-Potential Theory for Many-Valley Semiconductors with Anisotropic Scattering. *Physical Review*, 101(3):944–961, 1956.

[32] H. Kosina, M. Nedjalkov, and S. Selberherr. Monte Carlo Analysis of the Small-Signal Response of Charge Carriers. In *Large-Scale Scientific Computing*, pp 175–182, 2001.

[33] J. W. Harrison and J. R. Hauser. Alloy Scattering in Ternary III-V Compounds. *Physical Review B*, 13(12):5347–5350, 1976.

[34] H. Kosina and G. Kaiblinger-Grujin. Ionized-Impurity Scattering of Majority Electrons in Silicon. *Solid-State Electron.*, 42(3):331–338, 1998.

[35] H. Kosina. A Method to Reduce Small-Angle Scattering in Monte Carlo Device Analysis. *IEEE Trans.Electron Devices*, 46(6):1196–1200, 1999.

[36] M. V. Fischetti. Effect of the Electron-Plasmon Interaction on the Electron Mobility in Silicon. *Physical Review B*, 44(11):5527–5534, 1991.

[37] T. Kaneto. Calculation of Minority-Carrier Mobilities in Heavily Doped p-Type Semiconductors in the Dielectric-Function Formalism. *Physical Review B*, 47(24):16257–16266, 1993.

[38] M. Glicksman. Mobility of Electrons in Germanium-Silicon Alloys. *Physical Review*, 111(1):125–128, 1958.

[39] S. Inaba, M. Takahashi, Y. Okayama, A. Yagishita, F. MatsuokaAND, and H. Ishiuchi. Impact Of Trench Sidewall Interface Trap In Shallow Trench Isolation On Junction Leakage Current Characteristics For Sub-0.25 μm CMOS Devices. *Symp. on VLSI Technology*, pp 119–120, 1997.

[40] D. Ha, C. Cho, D. Shin, G. Koh, T. Chung, and K. Kim. Anomalous Junction Leakage Current Induced by STI Dislocations and its Impact on Dynamic Random Access Memory Devices. *IEEE Trans. Electron Devices*, 46(5):940–946, 1999.

[41] P. Smeys, P. Griffin, Z. Rek, I. De Wolf, and K. Saraswat. Influence of Process-Induced Stress on Device Characteristics and its Impact on Scaled Device Performance. *IEEE Trans. Electron Devices*, 46(6):1245–1252, 1999.

[42] V. Gopinath, H. Puchner, and M. Mirabedini. Impact of Shallow Trench Liner Oxidation Scheme on Junction Leakage. In H. Ryssel, G. Wachutka, and H. Grünbacher, editors, *31th European Solid-State Device Research Conference*, pp 419–422, Nuremberg, 2001. Frontier Group.

[43] V. Gopinath, H. Puchner, and M. Mirabedini. STI Stress Induced Increase in Reverse Bias Junction Capacitance. *IEEE Electron Device Lett.*, 2002.

[44] Avanti Corporation, Fremont, CA. *TSUPREM-4 Two-Dimensional Process Simulation Program, Version 2000.4*, 2000. http://www.avanticorp.com/products.

[45] V. Palankovski, N. Belova, T. Grasser, H. Puchner, S. Aronowitz, and S. Selberherr. A Methodology for Deep Sub-0.25 μm CMOS Technology Prediction. *IEEE Trans. Electron Devices*, 48(10):2331–2336, 2001.

[46] D. Rohrbacher. Statistical Treatment of Rejection of Deviant Values: Critical Values of Dixon's ”Q” Parameter and Related Subrange Ratios at the 95% Confidence Level. *Analytical Chemistry*, 63:139–146, 1991.