VISTA Status Report
December 2001

T. Grasser, M. Gritsch, C. Heitzinger,
A. Hössinger, S. Selberherr

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

Contents

1 Advanced Hybrid Cellular Based Approach for Three-Dimensional Etching and Deposition Simulation

1.1 Introduction

For three-dimensional simulation of etching and deposition processes cellular based algorithms [1, 2] have advantages to polygonal based algorithms and level set algorithms due to their high robustness. For instance the formation of voids which is a very serious problem for polygonal based algorithms is implicitly handled. Nevertheless the major drawback of the cellular based simulators is the cellular data format. On the one hand side the cellular resolution is not very high because the memory requirement increases dramatically by choosing a higher accuracy. But when a topography simulator is applied to front end process simulation also thin layers like the gate oxide in a MOS transistor have to be resolved accurately. This means that the cell size should be smaller than one tenth of the gate oxide thickness. For instance, if the volume of the simulation domain is 1 μm^3 and the gate oxide thickness is 5 nm 8 billon cells were necessary to discretize the simulation domain, while the error is still not less than 10%.

A way to reduce the memory requirement is to use more advanced discretization methods than a simple regular grid. A bucket octree based representation could be the method of choice, but it still suffers from the fact that the discrete representation is normally not compatible to the polygonal based data format of other simulators used for semiconductor process simulation. As a consequence, several conversions between the cellular and the polygonal data format are necessary, if the cellular based simulator is integrated into a process flow. Errors of the order of the resolution of the cellular representation are introduced by each conversion. As already mentioned, small structures (especially thin layers) present in the simulation domain may be seriously modified or even get lost in the worst case, although for instance, not even small modifications of the gate oxide thickness are tolerable.

We present an advanced hybrid cellular approach which attacks the discretization error problem from two sides. On the one hand side we avoid a discretization of the complete simulation domain by restricting the descritization to a layer around the surface attacked by the topography process and on the other hand side two levels of cell resolution are used. Both measures give a higher cellular resolution resulting in a smaller discretization error.

1.2 Internal Data Representations

During the simulation two internal data representations are used. The core of the topography simulator is based on a regular grid layed over the whole simulation domain. Each cell defined by this grid either can contain material information or another regular grid defining refined cells. Note that no physical memory is associated with this grid at the beginning. Just during the simulation the material information is generated and stored on grid cells.

The topology of the complete simulation domain is stored seperately in the so called Wafer State Server [3]. This Wafer State Server holds the structure of the simulation geometry in a thetrahedral volume mesh discretized format and provides functions to comfortably access information about the simulation domain. Thereby material cells used by the topography simulator can be easily generated during the simulation. To achieve an accaptable access time the mesh is stored on a finite octree inside the Wafer State Server.

1.3 Topography Process

First the upper surface (surface attacked by the topography process) of the simulation domain is extracted by the Wafer State Server and discretized pointwise.

Around each discretized point, refined topography simulator cells (called refined cell in the following) are generated. Each cell determines its material type by accessing information from the Wafer State Server. Refined cells containing the surface of the simulation domain become **surface cells** which contain additional information to perform the topography process, like the size of the structuring element related to the surface cells position.

After setting up the discretized surface the topography process is performed. Thereby all previously determined structuring elements are applied to the simulation domain. All cells covered by the smallest rectangular prisms containing the structuring elements are defined at that time. Defining a cell means that memory is allocated for the cell if necessary and all the informantion stored on the cell is prepared by the Wafer State Server. This procedure is indicated by the gray area in Fig. 1. All cells containing the surface, the front of the topography process and interfaces between different materials are additionally refined to further increase the resolution in such sensitive areas.

The simulation proceeds similar as in the classical cellular approach, meaning that the material type of all cells attacked by the topography process is changed during the simulation. The only differences to the classical approach are that

- memory is successively allocated whenever a cell is attacked outside of an allocated memory area.
- cells close to the process front (e.g. etch front) become refined cells.
- cells moving away from the process front are converted back to simple cells if they contain refined cells just of one material type.

The use of a two level cell resolution does not only save memory but also increases the performance of the simulation, because generally each cell has to be checked whether it is inside or outside of a structuring element in order to determine if the material type of a cell has to be changed. If a cell which is not refined is completely contained within or outside of the structuring element, its subcells need not to be checked. This significantly reduces the number of test operations, if the size of the unrefined cells is smaller than the size of the structuring element.

1.4 Post Processing

After completing the topography simulation the topology of the modified simulation domain has to be generated. Therefore a triangulated representation of the etchfront is determined from the cellular data structure, simply by generating two triangles from all interfaces between a material cell and vacuum cells. Since all cells around the front of the topography process are refined cells just triangles with half the area of one side of a refined cell are generated. In order to reduce the huge number of triangles and to improve the quality of the triangulated front two postprocessing routines are applied.

On the one hand side the manhatten like structure is smoothed by projecting each point into a plane with the minimum distance from points surrounding the point which has to be projected. By restricting the offset due to projection to the half length of the diagonal of a refined cell the generation of illegal structures (intersecting triangles) is avoided. Thereby the angle between triangles which should be coplanar is made quite small.

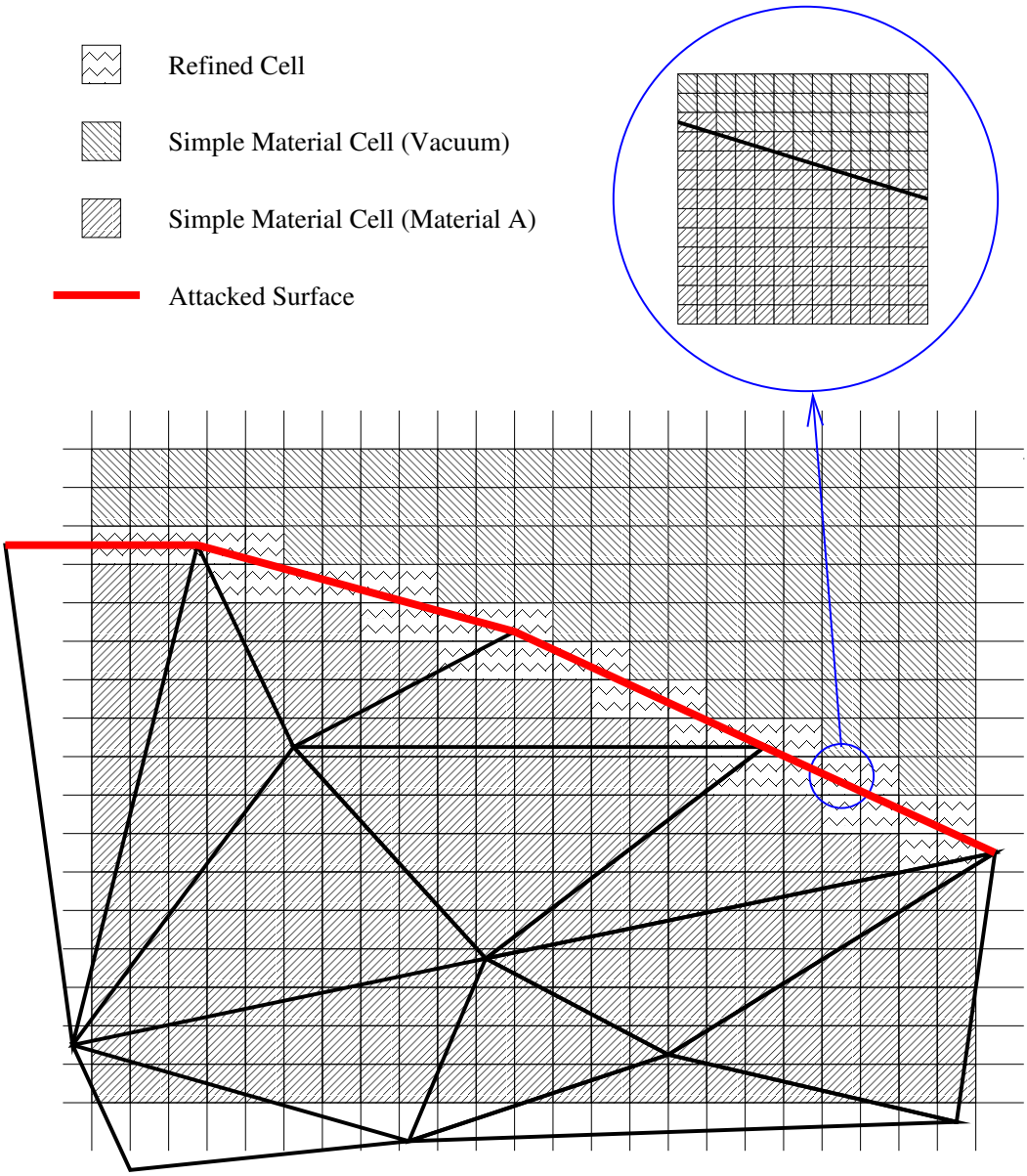

Figure 1: Illustration of the discretization procedure.

On the other hand side the number of triangles is reduced by removing edges [4] which are surrounded just by approximately coplanar triangles (the angle between the triangles is below a minimum tolerable angle). The number of triangles is usually significantly reduced and, additionally, more vestiges of the manhatten like structure are removed.

Finally a new wafer has to be set up which contains all the information available before the topography process, modified just by the effects of the topography process. At that point the major advantage of coupling the topgraphy simulator with the Wafer State Server shows up. During the simulation the Wafer State Server keeps track of mesh element attacked by the topography process. Thereby the Wafer State Server finally knows which segments have been attacked by the topography simulation and just these segments are modified. The modification is performed by constructing new segments from the old ones by cutting the segment surfaces with the etch front and remeshing the resulting topography. This procedure is schematically depicted in Fig. 2. In case of a deposition simulation no existing segments have

to be modified. Just new segments are created by meshing the space between the initial wafer surface and the final deposition fronts. Worth mentioning is that more than one front may have been generated during the simulation due to the formation of voids.

Fig. 3 shows a real three-dimensional example of the post processing flow for a highly unisotropic etching process. The brown silicon nitride layer is isotropically etched and the grey silicon dioxide layer acts as an etchstop. The selectivity of the etching process between silicon nitride and silicon dioxide was set to 7:1. In order to demonstrate that this approach can also be applied to deposition processes, as previously mentioned Fig. 4 and Fig. 5 show three-dimensional simulations of an isotropic and an unisotropic deposition process. In both cases a silicon trench was filled with silicon dioxide. For the unisotropic process the deposition aspect ratio was 0.15.

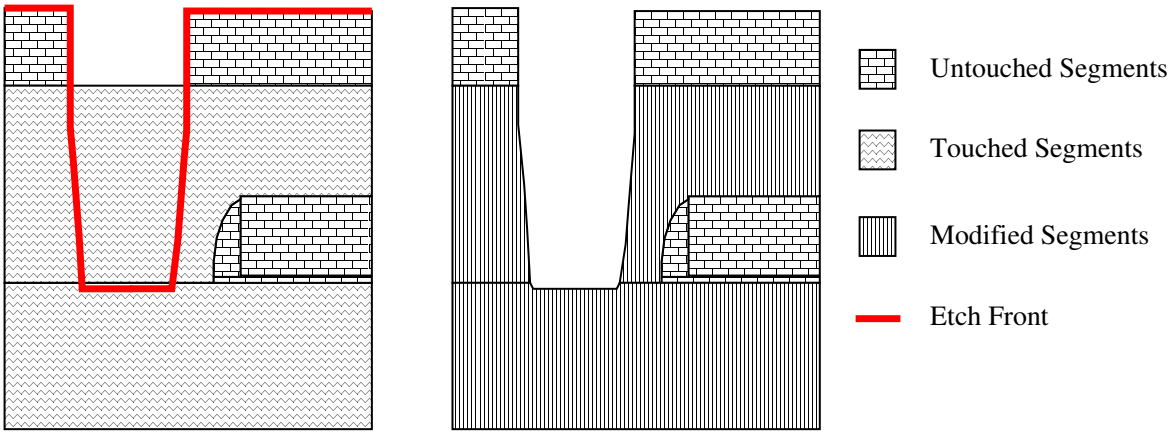

Figure 2: Illustration of the modification of the wafer state data due to the etch process.

1.5 Conclusion

We have presented a method for the simulation of etching and deposition processes which makes use of the robust cellular algorithm. The new method overcomes the problem of conversion errors when a cellular based simulator interacts with simulators based on polygonal data structures. Additionally the memory requirement is drastically reduced so that also large structures can be simulated with sufficiently high resolution.

Acknowledgment

This work is supported by LSI Logic, by Infineon, and by APART (Austrian Program for Advanced Research and Technology) from the Österreichische Akademie der Wissenschaften.

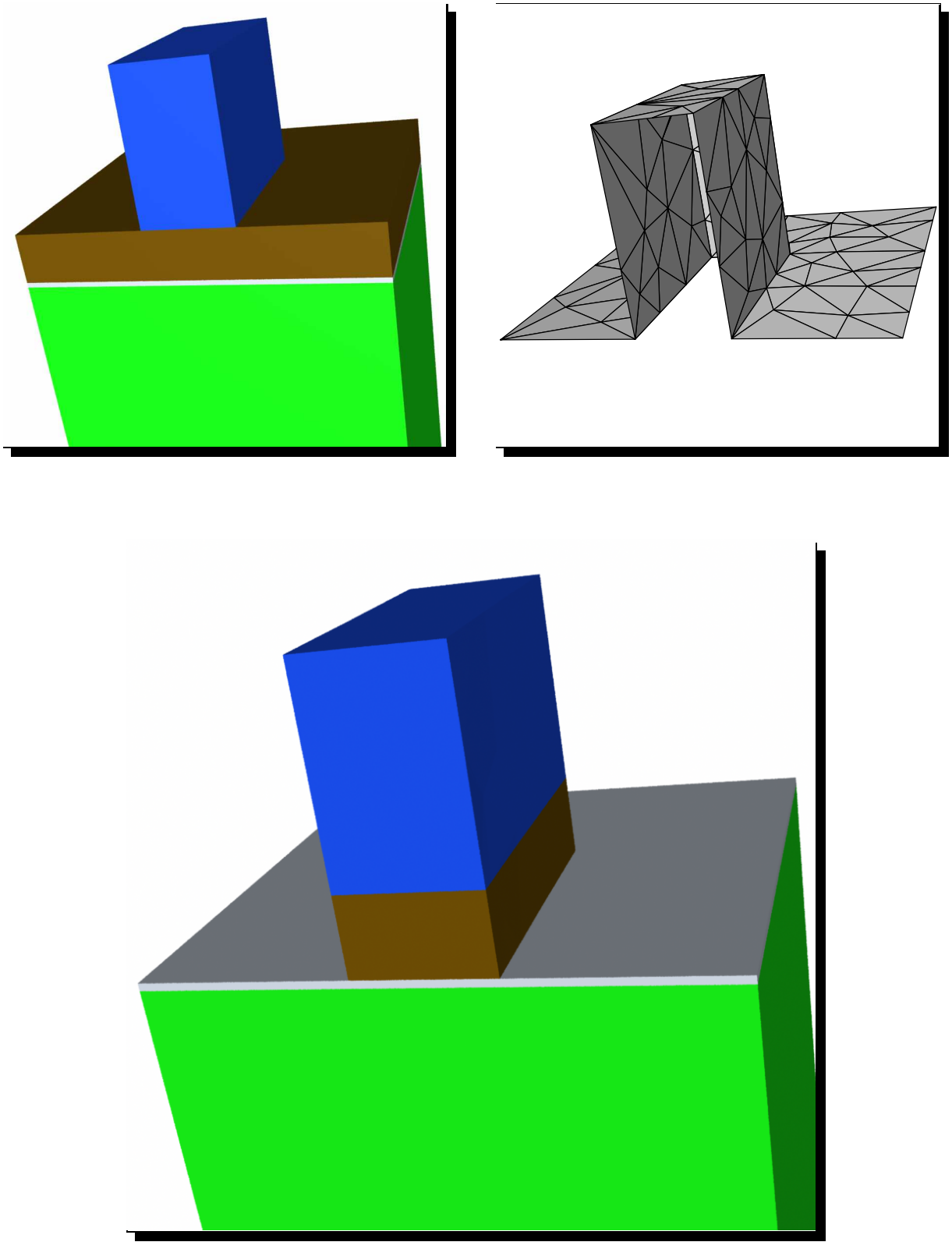

Figure 3: Post processing flow demonstrated for the case of a highly unisotropic etching process: Original structure (top-left), smoothed etch-front (top-right), final structure (bottom). The selectivity between resist and nitride was set to 0.3 and 0.15 between SiO_2 and nitride.

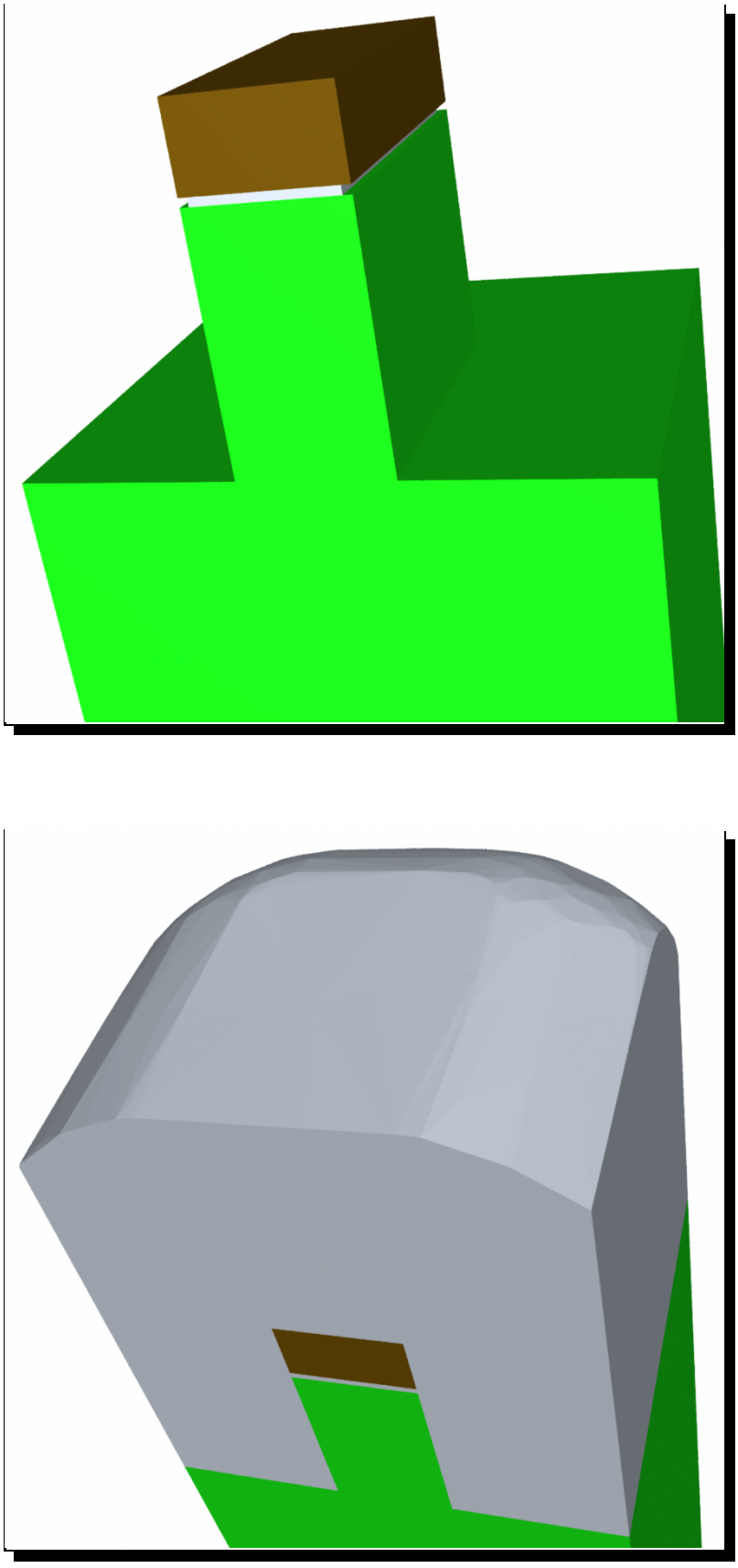

Figure 4: Three dimensional simulation of an isotropic deposition processes. The input structure (top) is a shallow silicon trench (300 nm). This trench is filled with silicon dioxide (bottom).

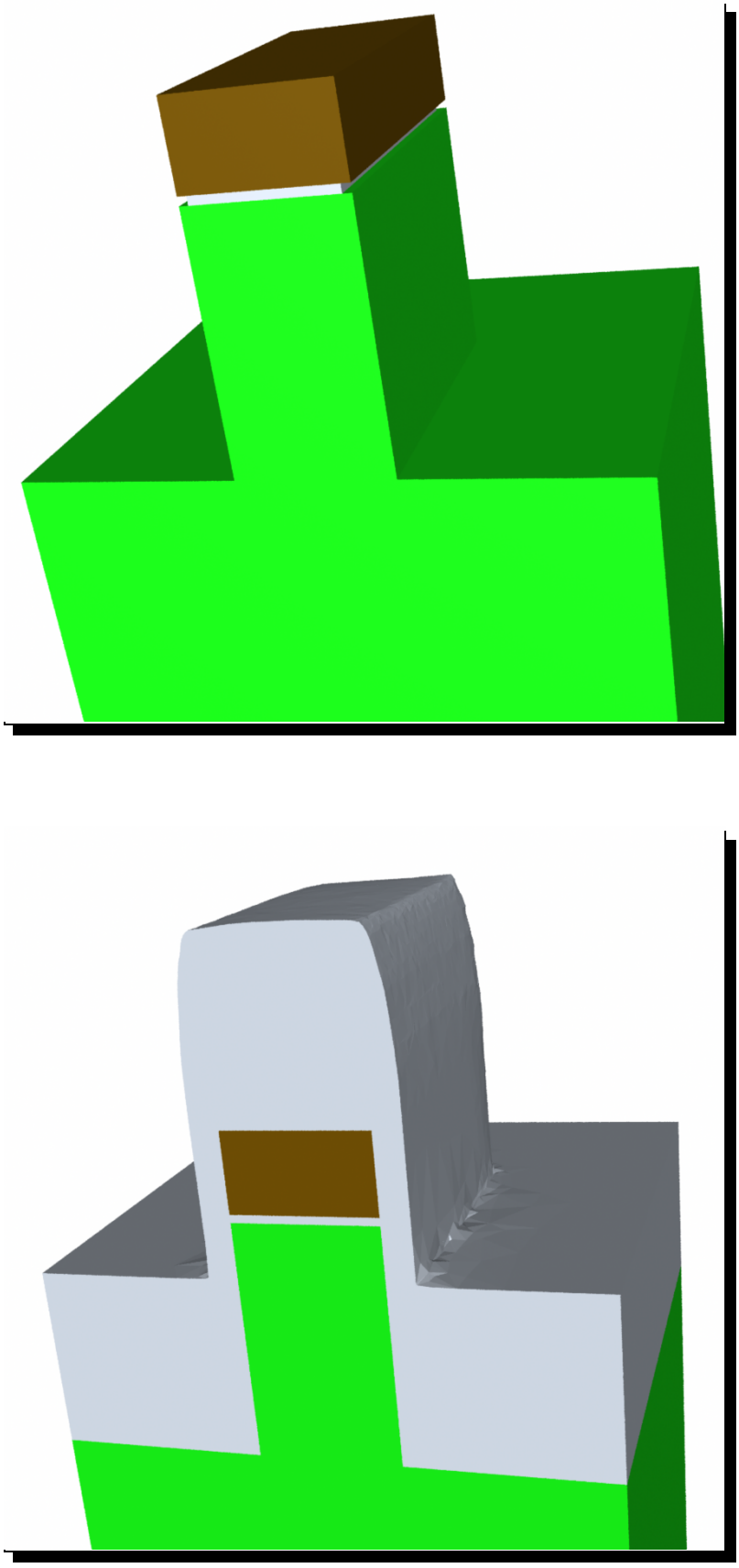

Figure 5: Three dimensional simulation of an isotropic deposition processes. The input structure (top) is a shallow silicon trench (300 nm). This trench is filled with silicon dioxide (bottom). The aspect ratio of the deposition was set to 0.15.

2 Simulation of Partially Depleted SOI MOSFETs using an Improved Hydrodynamic Transport Model

2.1 Introduction

The small minimum feature size of todays devices makes it more and more difficult to get proper simulation results using the widely accepted drift-diffusion (DD) transport model. In particular the lack of accounting for nonlocal effects such as carrier heating and velocity overshoot makes it desirable to use more sophisticated transport-models. These are obtained by considering the first three or four moments of the BOLTZMANN equation. However, these so called hydrodynamic models (HD) which are nowadays available in most of the device simulation programs, lead to interesting problems when applied to SOI MOSFETs.

2.2 Device Used

The simulations discussed in this paper were performed on a device with an assumed effective gate-length of $130\,\mathrm{nm}$, a gate-oxide thickness of $3\,\mathrm{nm}$, and a silicon-film thickness of $200\,\mathrm{nm}$. With a p-doping of $\mathrm{N_A} = 7.5 \cdot 10^{17}\ \mathrm{cm}^{-3}$ the device is partially depleted. The Gaussian-shaped n-doping under the electrodes has a maximum of $\mathrm{N_D} = 6 \cdot 10^{20}\ \mathrm{cm}^{-3}$.

2.3 Observed Behavior

By using the hydrodynamic transport model for simulation of the output characteristics of partially depleted SOI MOSFETs, an anomalous decrease of the drain current with increasing drain-source voltage can be observed [5, 6] (Fig. 6). The anomalous effect has been reproduced using two different device simulators, namely MINIMOS-NT [7] and DESSIS [8]. It is believed that this decrease is a spurious effect because to our knowledge it is neither present in experiments nor can it be observed when using the drift-diffusion (DD) transport model. One exception is given in [9], where a weak decrease of the drain-current is reported.

2.4 Cause of the Effect

The main difference between the HD and the DD transport model is given by the energy balance equation. The benefit of the increased computational effort is that the carrier temperature can differ from the lattice temperature. Since the diffusion of the carriers is proportional to their temperature, the diffusion can be significantly higher with the HD transport model. Fig. 7 clearly shows the enhanced vertical diffusion of electrons as compared with the DD result in Fig. 8.

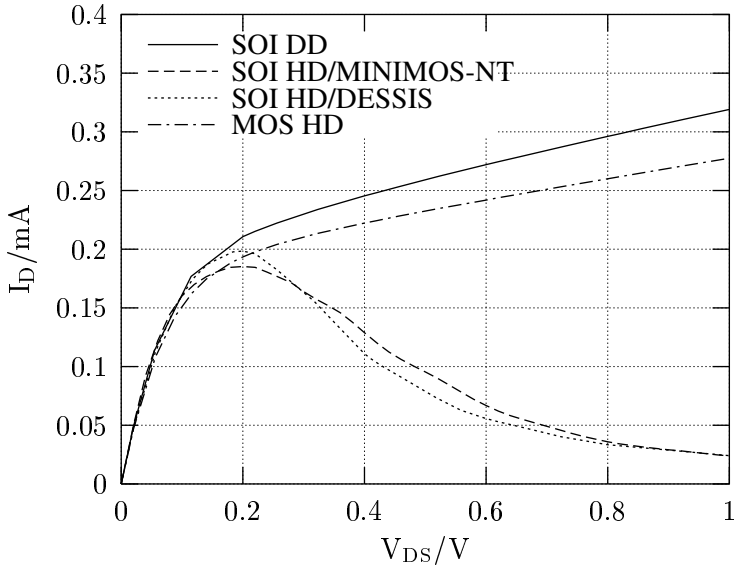

Figure 6: Output characteristics of the SOI obtained by DD an HD simulations using two different device simulators.

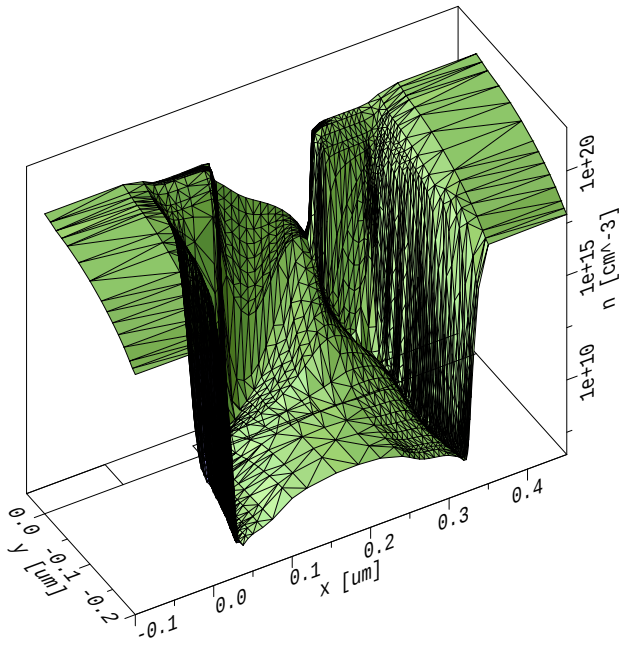

Figure 7: Electron concentration in an SOI MOSFET obtained by a HD simulation.

When simulating SOI MOSFETs this increased diffusion has a strong impact on the body potential, because the hot electrons of the pinch-off region have enough energy to overcome the energy barrier towards the floating body region and thus enter into the sea of holes. Some of these electrons in the floating body are sucked-off from the drain-body and source-body junctions, but most recombine. The holes removed by recombination cause the body potential to drop. A steady state is obtained when the body potential reaches a value which biases the junctions enough in reverse direction so that thermal generation of holes

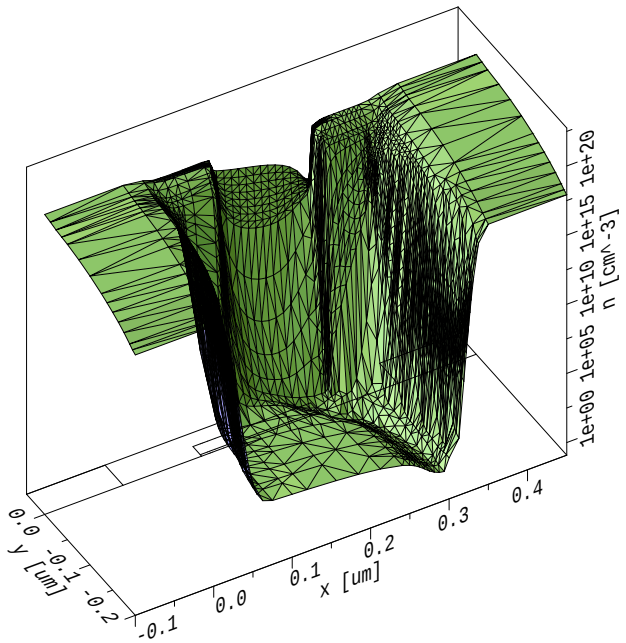

Figure 8: Electron concentration in an SOI MOSFET obtained by a DD simulation.

in the junctions can compensate this recombination process. The decrease in the output characteristics is directly connected to the drop of the body potential via the body-effect.

2.5 Weight Factors

Our first attempt to avoid the anomalous current decrease was to tune the empirical weight factors of thermal diffusion and heat flow, as provided by the HD model of DESSIS. Within this parameter-space only minor improvements in the IV characteristics were possible. Therefore, our investigations continued with more physically motivated modifications, using MINIMOS-NT.

2.6 Modifications

In Monte-Carlo (MC) simulations the spreading of hot carriers away from the interface is much less pronounced than in HD simulations. If we assume that the BOLTZMANN equation does not predict the hot carrier spreading, and if the HD equations derived from the BOLTZMANN equation do so, the problem must be introduced by the assumptions made in the derivation of the HD model. Relevant in this regard is the approximation of tensor quantities by scalars and the closure of the hierarchy of moment equations.

In order to capture more realistically the phenomenon of hot carrier diffusion we derived a HD equation set from the BOLTZMANN equation permitting an anisotropic temperature and a non-MAXWELLian distribution function. The current density $J_{n,\xi}$ and the energy density $S_{n,\xi}$ are given by

$$
\begin{aligned}
J_{n,\xi} &= \mu_n \left(\mathrm{k_B}\, \nabla_\xi \left(n\, T_{\xi\xi} \right) + \mathrm{q}\, E_\xi\, n \right) , \\
S_{n,\xi} &= -\frac{5}{2}\, \frac{\mathrm{k_B}}{\mathrm{q}}\, \mu_S \left(\mathrm{k_B}\, \nabla_\xi \left(n\, \beta_n\, T_{\xi\xi}\, \Theta \right) + \mathrm{q}\, E_\xi\, n\, \Theta \right) , \\
&\text{with} \quad \nabla_\xi = \frac{\partial}{\partial \xi} \quad \text{and} \quad \Theta = \frac{3\, T_n + 2\, T_{\xi\xi}}{5} .
\end{aligned}
$$

$T_{\xi\xi}$ denotes the diagonal component of the temperature tensor for direction $\vec{e}_\xi$. Off-diagonal components are neglected. β_n is the normalized moment of fourth order. By setting $T_{\xi\xi} = T_n$ and $\beta_n = 1$ the conventional HD model is obtained. The solution variable is still the carrier temperature T_n, whereas the tensor components and the fourth moment are modeled empirically as functions of the carrier temperature. First empirical modeling of $T_{\xi\xi}$ was performed by distinguishing between directions parallel and normal to the current density:

$$T_{\xi\xi} = T_{xx}\,\cos^2\varphi + T_{yy}\,\sin^2\varphi\,, \quad T_{xx} = \gamma_x T_n\,, \quad T_{yy} = \gamma_y T_n\,,$$

$$\gamma_\nu(T_n) = \gamma_{0\nu} + \left(1-\gamma_{0\nu}\right)\,\exp\!\left(-\left(\frac{T_n - T_L}{\mathrm{T}_{\mathrm{ref},\gamma}}\right)^2\right)\,, \quad \nu = x, y\,.$$

The anisotropy functions $\gamma_\nu(T_n)$ assume 1 for $T_n = T_L$ and an asymptotic value $\gamma_{0\nu}$ for large T_n, ensuring that only for sufficiently hot carriers the distribution becomes anisotropic, whereas the equilibrium distribution stays isotropic. With respect to numerical stability the transition should not be too steep. $\mathrm{T}_{\mathrm{ref},\gamma} = 600\,\mathrm{K}$ appeared to be appropriate.

Another effect observed in MC simulations is that in most parts of the channel the high energy tail is less populated than that of a MAXWELLian distribution, which gives $\beta_n < 1$ (Fig. 11). A simple model for β_n was used.

$$\beta_n(T_n) = \beta_0 + \left(1-\beta_0\right)\,\exp\!\left(-\left(\frac{T_n - T_L}{\mathrm{T}_{\mathrm{ref},\beta}}\right)^2\right)$$

Again, this expression ensures that only for sufficiently large T_n the distribution deviates from the MAXWELLian shape.

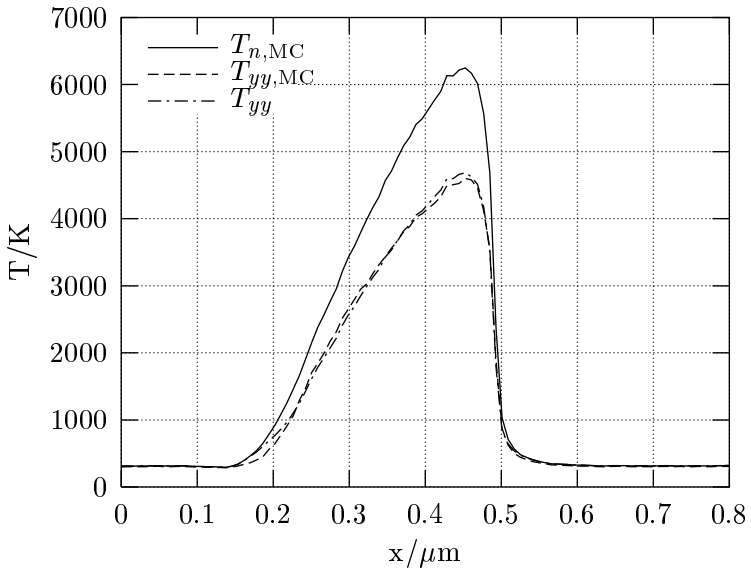

Figure 9: MC simulation of an nin-structure showing the x-component of the temperature compared to the mean temperature $T_{n,\mathrm{MC}}$. The analytical T_{yy} uses $\gamma_{0\nu} = 0.75$.

2.7 Results

The modified flux equations have been implemented in MINIMOS-NT using a straight forward extension of the Scharfetter-Gummel discretization scheme. Numerical stability does not degrade as compared to standard HD simulations. Parameter values were estimated from MC results for one-dimensional test

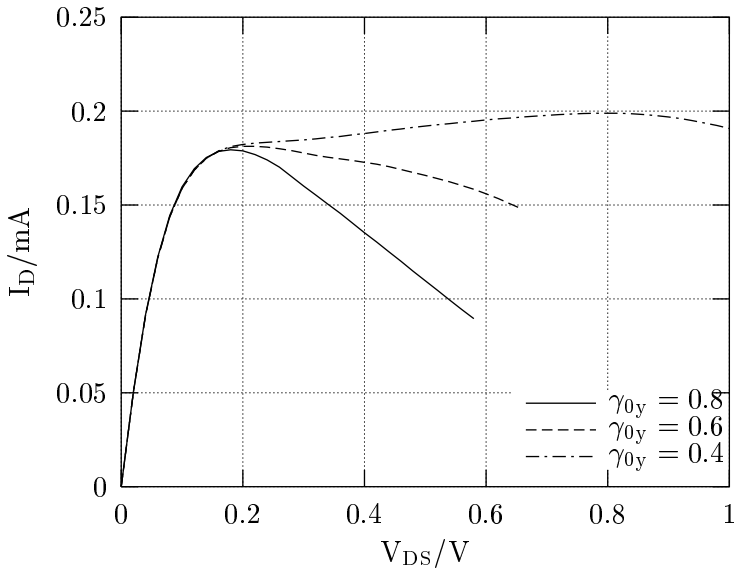

Figure 10: Output characteristics of the SOI obtained by anisotropic HD simulations.

structures. Fig. 9 indicates that $\gamma_{0y} = 0.75$ is a realistic value for the anisotropy parameter. Fig. 10 shows the influence of γ_{0y} on the output characteristics. By accounting for a reduced vertical temperature it is possible to reduce the spurious current decrease, but only to a certain degree and by assuming a fairly large anisotropy. MC simulations yield values close to $\beta_0 = 0.75$ for the non-MAXWELLian parameter in the channel region (Fig. 11). This parameter shows only a weak dependence on doping and applied voltage.

By combining the modifications for an anisotropic temperature and a non-MAXWELLian closure relation the artificial current decrease gets eliminated (Fig. 12). Parameter values roughly estimated from MC simulations can be used, e.g. $\gamma_{0y} = 0.75$ and $\beta_0 = 0.75$. In the parameter range where the current drop is eliminated the output characteristics are found to be rather insensitive to the exact parameter values.

2.8 Conclusion

Standard HD simulations of SOI MOSFET give anomalous output characteristics. To solve this problem, an improved HD transport model has beed developed. By including two distinct modifications, namely an anisotropic carrier temperature and a modified closure relation, the spurious diffusion of hot electrons in the vertical direction has been sufficiently reduced. Further careful modeling of these two effects on the basis of MC data may be required.

Acknowledgement

This work has been supported by *Intel Corp.*, Santa Clara, and the *Christian Doppler Gesellschaft*, Vienna.

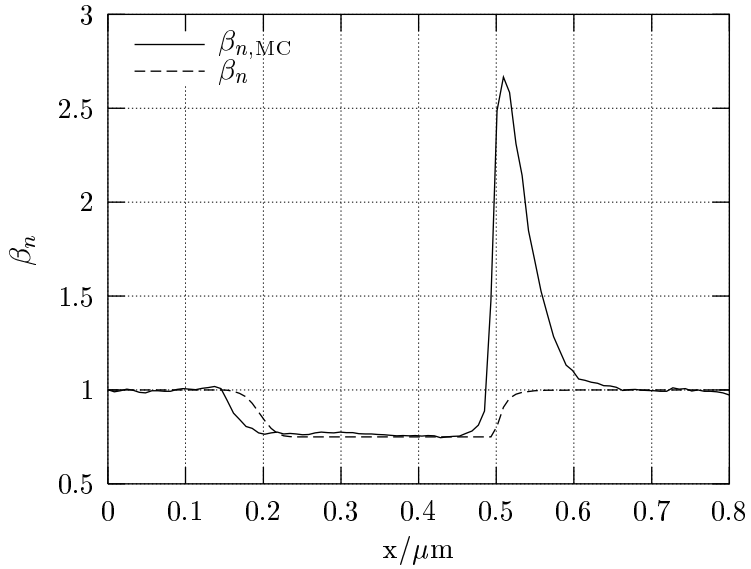

Figure 11: MC simulation of a nin-structure showing the normalized moment of fourth order $\beta_{n,\mathrm{MC}}$ compared to the analytical β_n with $\beta_0 = 0.75$.

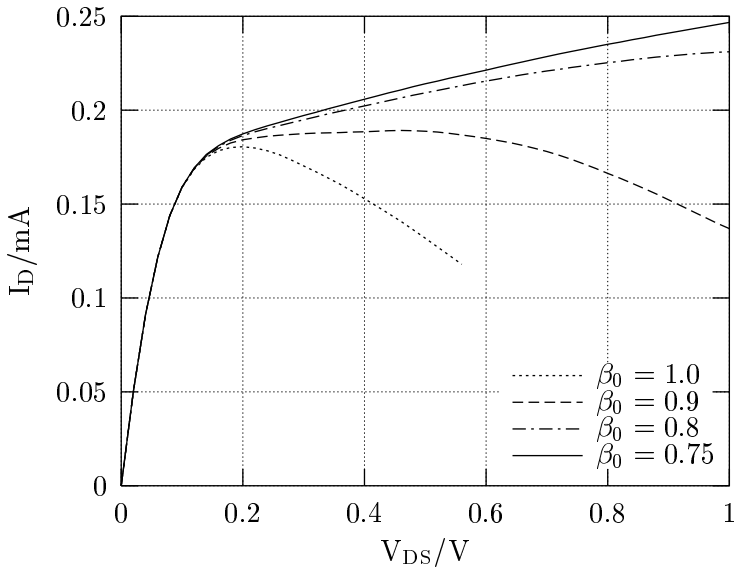

Figure 12: Output characteristics of the SOI assuming an anisotropic temperature ($\gamma_{0y} = 0.75$) and a modified closure relation.

3 A Calibrated Model for Silicon Self-Interstitial Cluster Formation and Dissolution

3.1 Introduction

The formation and dissolution of Silicon self-interstitial clusters is linked to the phenomenon of TED (transient enhanced diffusion), which in turn has gained importance in the manufacturing of semiconductor devices. Based on theoretical considerations and measurements of the number of self-interstitial clusters during a thermal step we were interested in finding a suitable model for the formation and dissolution of Silicon self-interstitial clusters and extracting corresponding model parameters for two different technologies (i.e., material parameter sets). In order to automate the inverse modeling part a general optimization framework was used. Additional to solving this problem the same setup can solve a wide range of inverse modeling problems occuring in the domain of process simulation.

The goal is to find a calibrated model for the formation and dissolution of Silicon self-interstitial clusters of $\{113\}$ or $\{311\}$ defects. Finding a good calibrated model for self-interstitial clustering is important for accurately simulating the TED (transient enhanced diffusion) of impurities, which is influenced by these self-interstitial clusters. TED is the fast displacement of impurities in the first thermal step just after implantation and the simulation of its evolution and magnitude is important in the manufacturing processes of submicron devices.

The source of of the Silicon self-interstitials was shown to be the $\{113\}$ defects, which are rod like clusters of interstitials [10]. Counting the amount of self-interstitials is a non-trivial task: from transmission electron micrographs the number of interstitials in each defect and thus the total number has to be measured. In [11] one can find measurements giving the number of interstitials as a function of time for annealing at four temperatures (670°C, 705°C, 738°C, and 815°C) and $5 \cdot 10^{13}\mathrm{cm}^{-2}$, 40keV implants. These measurements are shown in detail in Figure 13 and provided the basis for this inverse modeling problem. For the computations we used TSUPREM4 [12] and the optimization framework SIESTA [13, 14].

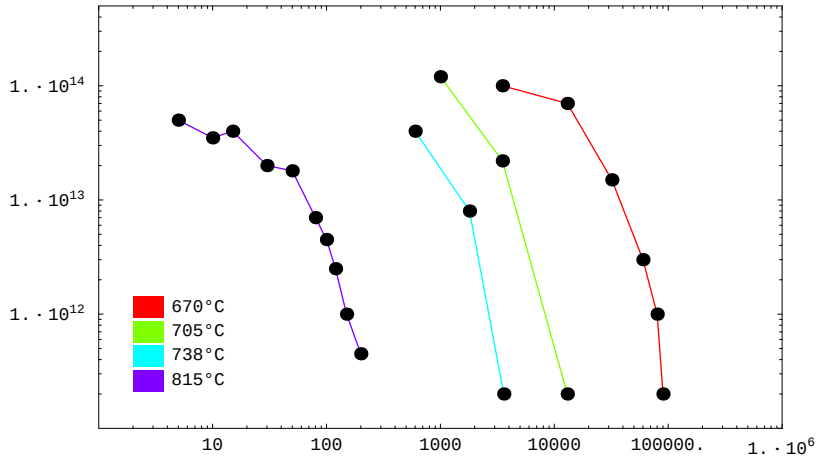

Figure 13: The Silicon self-interstitial density (in cm^{-3}) as a function of time (in s) for different annealing temperatures (interstitials stored in $\{113\}$ or $\{311\}$ defects after $5 \cdot 10^{13}\mathrm{cm}^{-2}$, 40keV implants).

Variable	Interval	Unit
d0	$[25, 1000]$	$\mathrm{cm}^2\mathrm{s}^{-1}$
dE	$[1.4, 1.85]$	eV
kfi0	$[10^{20}, 10^{28}]$	$\mathrm{cm}^{-3(1+\mathrm{isfi}-\mathrm{ifi})}\mathrm{s}^{-1}$
kfiE	$[3.4, 6.0]$	eV
ifi	constant $= 2$	1
isfi	constant $= 2$	1
kfc0	$[10^{17}, 7 \cdot 10^{19}]$	$\mathrm{cm}^{-3(1+\mathrm{isfc}-\mathrm{ifc}-\mathrm{cf})}\mathrm{s}^{-1}$
kfcE	$[4.9, 5.2]$	eV
ifc	constant $= 1$	1
isfc	constant $= 1$	1
α	$[0, 5000]$	1
cf	constant $= 1$	1
kr0	$[1.5 \cdot 10^{16}, 10^{18}]$	$\mathrm{cm}^{-3(1-\mathrm{cr})}\mathrm{s}^{-1}$
krE	$[3.0, 3.62]$	eV
cr	constant $= 1$	1

Table 1: Variables, their intervals, and their units.

We were interested in finding solutions for two different technologies corresponding to different values of several TSUPREM4 variables. In the following we will call these parameter sets the high and the low parameter set (the latter being the TSUPREM4 default values).

Since the rate of formation and dissolution is not yet fully understood, the model used contains several proposed models (e.g., [15]) as special cases [12]. After describing the model and the details of the inverse modeling problem we present the results and the calibrated model.

3.2 Modeling Silicon Self-Interstitial Cluster Formation and Dissolution

In [15] the following equation describing interstitial cluster kinetics is given:

$$\frac{\partial C}{\partial t} = 4\pi\alpha a D_I I C - \frac{C D_I}{a^2} \mathrm{e}^{-E_b/kT}, \tag{1}$$

where $D_I = D_0 \mathrm{e}^{-E_m/kT}$ is the interstitial diffusivity, a is the average interatomic spacing, α is the capture radius expressed in units of a, $C(t, x)$ is the concentration of interstitials trapped in clusters, $I(t, x)$ is the concentration of free interstitials and T is the annealing temperature.

Here the main formula of the model for the change of the concentration of clustered interstitials is

$$\frac{\partial C}{\partial t} = K_{\mathrm{fi}} \frac{I^{\mathrm{ifi}}}{I_*^{\mathrm{isfi}}} + K_{\mathrm{fc}} \frac{I^{\mathrm{ifc}}}{I_*^{\mathrm{isfc}}} (C + \alpha I)^{\mathrm{cf}} - K_r C^{\mathrm{cr}}, \tag{2}$$

where $C(t, x)$ denotes the concentration of clustered interstitials, t time, $I(t, x)$ the concentration of unclustered interstitials, and $I_*(t, x)$ the equilibrium concentration of interstitials (which can be found by solving $\partial C(t, x)/\partial t = 0$). There is a number of parameters to be adjusted: K_{fi}, K_{fc}, K_{r} (the reaction constants); the exponents I^{ifi}, I^{isfi}, I^{ifc}, and I^{isfc}, cf, and cr; and finally α.

The reaction constants have the form

$$\begin{aligned} K_{\mathrm{fi}} &= \mathrm{kfi0} \cdot \mathrm{e}^{-\mathrm{kfiE}/kT}, \\ K_{\mathrm{fc}} &= \mathrm{kfc0} \cdot \mathrm{e}^{-\mathrm{kfcE}/kT}, \\ K_{\mathrm{r}} &= \mathrm{kr0} \cdot \mathrm{e}^{-\mathrm{krE}/kT}, \end{aligned}$$

High parameter set (mean relative error 0.389666):

Variable	SIESTA variable	Best point found
$\mathrm{d0}$	d-0	51.7282
dE	d-e	1.76996
$\mathrm{kfi0}$	kfi-0	$4.97576 \cdot 10^{24}$
kfiE	kfi-e	3.77408
$\mathrm{kfc0}$	kfc-0	$4.36789 \cdot 10^{19}$
kfcE	kfc-e	4.95
α	kfci	1099.63
$\mathrm{kr0}$	kr-0	$2.77935 \cdot 10^{16}$
krE	kr-e	3.56997

Low parameter set (mean relative error 0.504462):

Variable	SIESTA variable	Best point found
$\mathrm{kfi0}$	kfi-0	$1.14156 \cdot 10^{25}$
kfiE	kfi-e	3.94079
$\mathrm{kfc0}$	kfc-0	$1.5051 \cdot 10^{19}$
kfcE	kfc-e	5.81858
α	kfci	1563.1
cf	cf	1.01287
$\mathrm{kr0}$	kr-0	$1.06467 \cdot 10^{17}$
krE	kr-e	3.84503
cr	cr	0.9639

Table 2: Results for the high and low parameter sets with the above free variables. The mean relative error found is 0.389666 for the high parameter set and 0.504462 for the low parameter set.

with $\mathrm{kfi0} > 0$, $\mathrm{kfc0} > 0$, and $\mathrm{kr0} > 0$. Here T is the temperature (in Kelvin) and $k = 8.617 \cdot 10^{-5} \mathrm{eV} \cdot \mathrm{K}^{-1}$ the Boltzmann constant. Since the coefficients $\mathrm{kfi0}$, $\mathrm{kfc0}$, and $\mathrm{kr0}$ are positive, the first two terms in (2) are responsible for the formation of clusters and the last term for the dissolution. The sum of interstitials counted in C and I remains constant and the intial value of C is $10^9 \mathrm{cm}^{-3}$.

The ratio I/I_* of the concentration of the unclustered interstitials and its equilibrium concentration is often called the interstitial supersaturation. Here additional exponents modify the interstitial supersaturation which appears in the form $I^{\mathrm{ifi}}/I_*^{\mathrm{isfi}}$ and $I^{\mathrm{ifc}}/I_*^{\mathrm{isfc}}$.

The first term $K_{\mathrm{fi}}(I^{\mathrm{ifi}}/I_*^{\mathrm{isfi}})$ describes the joining of two clusters and thus the expected values for the exponents are $\mathrm{ifi} = 2 = \mathrm{isfi}$. The second growth term $K_{\mathrm{fc}}(I^{\mathrm{ifc}}/I_*^{\mathrm{isfc}})(C + \alpha I)^{\mathrm{cf}}$ governs the case when an unclustered interstitials joins an interstitial cluster. Here we can expect the exponents to be 1. The second factor is a linear combination of C and I with an exponent.

Comparing (1) and (2), the growth term of (1), basically being a reaction constant times IC, is split into two parts providing greater flexibility: one depending on a modified interstitial supersaturation term and one depending on a modified interstitial supersaturation term times $(C + \alpha I)^{\mathrm{cf}}$. In the dissolution term an exponent, which was later found to be 1, is added.

3.3 Inverse Modeling and Results

Two points from the measurements in [11] were ignored since they were above the implanted dose. All measurements were viewed as one vector m. Let s be the vector of simulation results depending on the

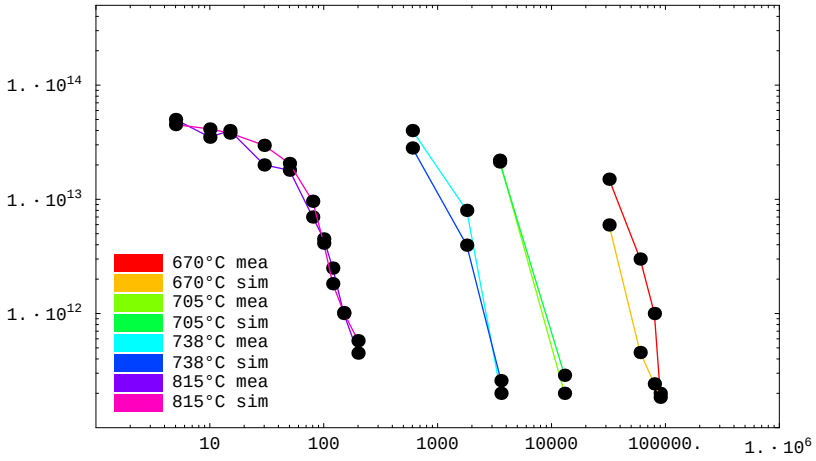

Figure 14: Result for the high parameter set corresponding to parameter values shown in Table 2. The logarithm (base 10) of the simulated and measured concentration $[\mathrm{cm}^{-3}]$ of interstitial clusters is shown depending on time $[\mathrm{s}]$.

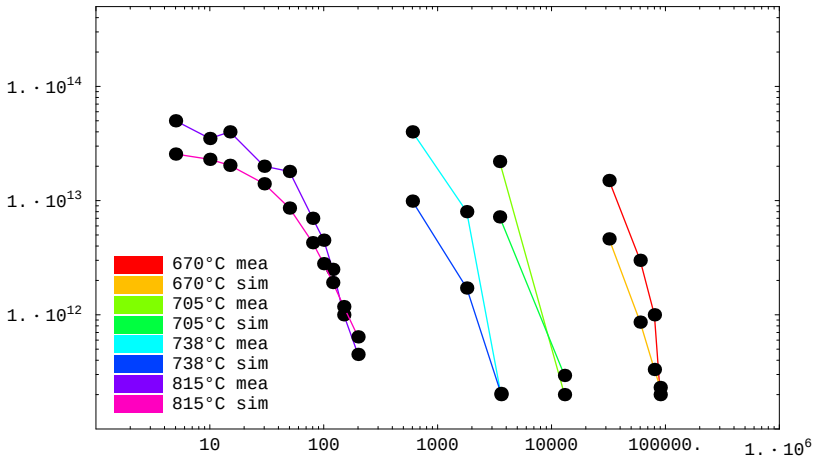

Figure 15: Result for the low parameter set corresponding to parameter values shown in Table 2. The logarithm (base 10) of the simulated and measured concentration $[\mathrm{cm}^{-3}]$ of interstitial clusters is shown depending on time $[\mathrm{s}]$.

parameters p to be identified. The objective function $f(p)$ to be minimized was the quadratic mean of the elementwise relative error between a simulated point and a measured point, i.e.,

$$f(p) := \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\frac{s_k(p)-m_k}{m_k}\right)^2}.$$

The variables of the objective function $f(p)$ are shown in Table 1. The variables $\mathrm{d0}$ and dE determine the diffusivity $\mathrm{d0}\cdot \mathrm{e}^{-\mathrm{dE}/kT}$ of interstitials in Silicon.

In order to reduce the time needed for the inverse modeling task, the optimization framework SIESTA was used. Its main tasks are optimizing a given objective function and parallelizing the executions of the

objective function which usually entails calling simulation tools in a loosely connected cluster of workstations. SIESTA provides several local and global optimizers, the ability to define complicated objective functions, and finally an interface to MATHEMATICA for examining the results.

The optimization approach was to first identify reasonable ranges for the variables with great influence, namely energies and exponents. While identifying these ranges suitable starting points for gradient based optimization were found as well. Using these ranges and starting points the optimizations proceeded automatically including all variables.

It was soon found that changing cf and cr did not yield improvements and whenever these variables could be used by a gradient based optimizer, values very close to 1 resulted. The results described in Table 2 and Figure 14 were obtained for the high parameter set.

Similarly we carried out the same computations for the low parameter set, i.e., TSUPREM4's default values. These results are shown in Table 2 and Figure 15.

3.4 Conclusion

Starting with (2) and the measurements from [11] we adjusted a model for the formation and dissolution of Silicon self-interstitial clusters, namely

$$\frac{\partial C}{\partial t} = K_{\mathrm{fi}}\frac{I^2}{I_*^2} + K_{\mathrm{fc}}\frac{I}{I_*}(C + \alpha I) - K_r C,$$

with values from Table 2. Although different values were also examined, the exponents in the first term were found to be equal to 2 ($\mathrm{ifi} = 2 = \mathrm{isfi}$), because two isolated interstitials can form a new cluster. Good results were achieved with $\mathrm{cf} = 1$ and $\mathrm{ifc} = 1 = \mathrm{isfc}$, which means the rate of free interstitials joining already existing clusters depends linearly on the number of excess interstitials (interstitials above the equilibrium concentration) and a linear combination of the number of clusters and interstitials. Finally the exponent cr was found to be 1. This means that the rate of dissolution depends on the concentration of clustered interstitials and on the factor K_r.

The terms responsible for cluster formation in both models don't share a common structure, thus we finally compare the results for the dissolution term. In (1) the dissolution term is

$$-\frac{D_0 C}{a^2}\mathrm{e}^{-(E_b+E_m)/kT},$$

where the values for E_b and E_m given in [15] are $E_b = 1.8\mathrm{eV}$ and $E_m = 1.77\mathrm{eV}$. $E_b + E_m = 3.57\mathrm{eV}$ agrees very well with the values found for krE in Table 2, namely $3.56997\mathrm{eV}$ for the high parameter set and $3.84503\mathrm{eV}$ for the low parameter set.

In order to give a summary, a refined model for the formation and dissolution of Silicon self-interstitial clusters was calibrated to published measurements for two different technologies (corresponding to two different sets of material parameters) and very good agreement was achieved.

Acknowledgement

This work has been supported by Sony Corporation, Atsugi, Japan.

4 An Accurate Impact Ionization Model which Accounts for Hot and Cold Carrier Populations

4.1 Introduction

Conventional macroscopic impact ionization models which use the average carrier energy as main parameter cannot accurately describe the phenomenon in modern miniaturized devices. Here we present a new model which is based on an analytic expression for the distribution function. In particular, the distribution function model accounts explicitly for a hot and a cold carrier population in the drain region of MOS transistors. The parameters are determined by three even moments obtained from a solution of a six moments transport model. Together with a nonparabolic description of the density of states accurate closed form macroscopic impact ionization models can be derived based on familiar microscopic descriptions.

Accurate calculation of impact ionization rates in macroscopic transport models is becoming more and more important due to the ongoing feature size reduction of modern semiconductor devices. Conventional models which use the average carrier energy as main parameter fail because impact ionization is very sensitive to the shape of the distribution function, in particular to the high-energy tail. The average energy is not sufficient for obtaining this information and higher order moments of the distribution function have to be considered. We favor a local description because the scattering operator in Boltzmann's equation is a functional of the local distribution function which should be reflected by the model.

4.2 The Refined Model

Here we present a refined version of a previously published model [16] developed for the use with a six moments transport model [17] which also accounts for the kurtosis of the distribution function, $\beta_n = (3/5)\langle\mathcal{E}^2\rangle/\langle\mathcal{E}\rangle^2$, in addition to the carrier temperature. Sofar, we used

$$f(\mathcal{E}) = A\exp\Big[-\Big(\frac{\mathcal{E}}{\mathcal{E}_{\rm ref}}\Big)^b\Big] \tag{3}$$

$$g(\mathcal{E}) = g_0\sqrt{\mathcal{E}}\Big(1+(\eta\mathcal{E})^\zeta\Big) \tag{4}$$

for the symmetric part of the distribution function and for the nonparabolic density of states, respectively. The parameters $\mathcal{E}_{\rm ref} = \mathcal{E}_{\rm ref}(T_n,\beta_n)$ and $b = b(T_n,\beta_n)$ are functions of the carrier temperature and kurtosis and are determined in such a way that $f(\mathcal{E})$ reproduces the given moments T_n and β_n [16]. The parameters η and ζ of (4) are determined by a fit to either Kane's dispersion relation [18] or to pseudo-potential data [16]. Due to the form of the fit expression for the density of states (4) it is possible to give algebraic expressions for the moments of (3) using Gamma functions. With the parameter values $\eta = 1.4\,\mathrm{eV}^{-1}$ and $\zeta = 1.08$ the error in the first three even moments was found to be smaller than 1% when compared to the results obtained from Kane's relation.

Expression (3) is accurate inside the channel of MOS transistors and gives values of $b > 1$. Although (3) gives reasonable approximations for the distribution function inside the drain region ($b < 1$) there are two problems: Firstly, the high-energy tail is overestimated and secondly, during the transition from the channel to the drain region, the exponent b assumes the value 1 which corresponds to a Maxwellian distribution function, a result not confirmed by Monte Carlo simulations. This error is amplified when for instance impact ionization rates are calculated where only the high-energy tail of the distribution function is required [16].

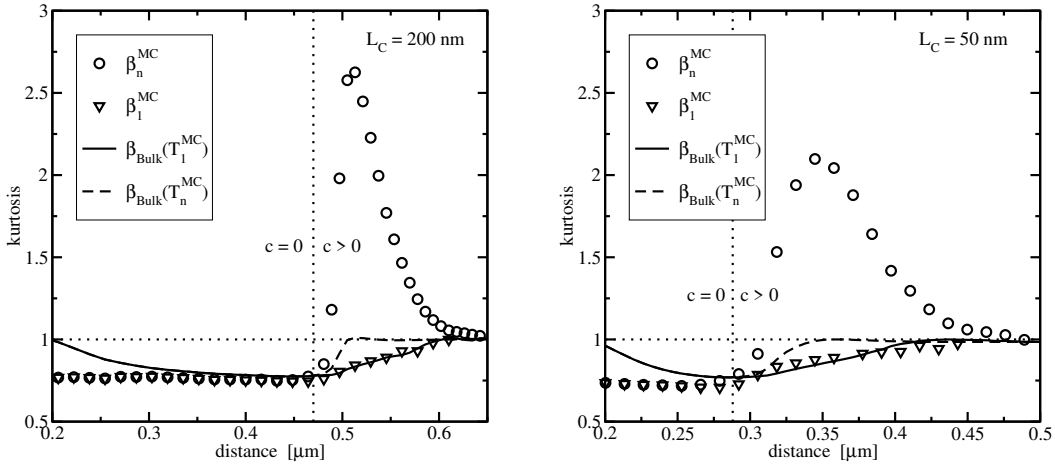

Figure 16: Comparison of β_1^{MC} with two analytical models. When the temperature of the hot distribution function T_1 is used in the bulk characteristic, accurate results are obtained for n^+-n-n^+ test-structures with $L_{\mathrm{C}} = 200$ nm and $L_{\mathrm{C}} = 50$ nm. Note that $\beta_{\mathrm{Bulk}}(T_n)$ does not properly describe the behavior of β_1. Also shown is the kurtosis of the total distribution function β_n.

To improve the model we note that at the drain junction the hot carriers from the channel meet a large pool of cold carriers and two populations coexist. We account for this fact by using a superposition of two distributions, similar to the work of Sonoda *et al.* [19]

$$f(\mathcal{E}) = A\Big\{\underbrace{\exp\Big[-\Big(\frac{\mathcal{E}}{\mathcal{E}_{\mathrm{ref}}}\Big)^b\Big]}_{f_1(\mathcal{E})} + c\,\underbrace{\exp\Big[-\frac{\mathcal{E}}{\mathrm{k_B}T_2}\Big]}_{f_2(\mathcal{E})}\Big\} \tag{5}$$

We now have to determine the five parameters A, $\mathcal{E}_{\mathrm{ref}}$, b, c, and T_2 which describe the distribution function, that is, we need two heuristic relations in addition to the three parameters n, T_n, and β_n provided by the six moments model. To get an idea about the behavior of the distribution function in the drain region we look at the second order moment T_1 and the fourth order moment β_1 of the hot distribution $f_1(\mathcal{E})$ only (Fig. 16) which was extracted from the total distribution function obtained by a Monte Carlo simulation in a post-processing step. Additional simulations show that the temperature T_2 of the cold Maxwellian distribution function $f_2(\mathcal{E})$ rapidly relaxes to the lattice temperature T_{L} and will be modeled as $T_2 = T_{\mathrm{L}}$ in this work. The kurtosis β_1, however, is crucial for an accurate description of the high-energy tail. Interestingly, β_1 can be modeled accurately via the bulk relation $\beta_{\mathrm{Bulk}}(T_n)$ which can be derived from the homogeneous six moments model [17] as

$$\beta_{\mathrm{Bulk}}(T_n) = \frac{T_{\mathrm{L}}^2}{T_n^2} + 2\,\frac{\tau_\beta}{\tau_\mathcal{E}}\frac{\mu_S}{\mu_n}\Big(1 - \frac{T_{\mathrm{L}}}{T_n}\Big) \tag{6}$$

where $\tau_\mathcal{E}$, τ_β, μ_n, and μ_S are the energy relaxation time, the kurtosis relaxation time, the electron mobility, and the energy flux mobility, respectively.

A comparison of the model $\beta_1 = \beta_{\mathrm{Bulk}}(T_1)$ with Monte Carlo data is shown in Fig. 16. where the temperature of the high-energy tail T_1 has been taken as the argument. Note that $\beta_{\mathrm{Bulk}}(T_n)$ approaches unity too quickly as also shown in Fig. 16 which underlines the idea of modeling the hot and cold electrons as separate populations.

For the calculation of the parameters $\mathcal{E}_{\mathrm{ref}}$, b, and c we have to detect the regions where a cold population exists. Monte Carlo simulations show that inside the channel the tail of the distribution function is always

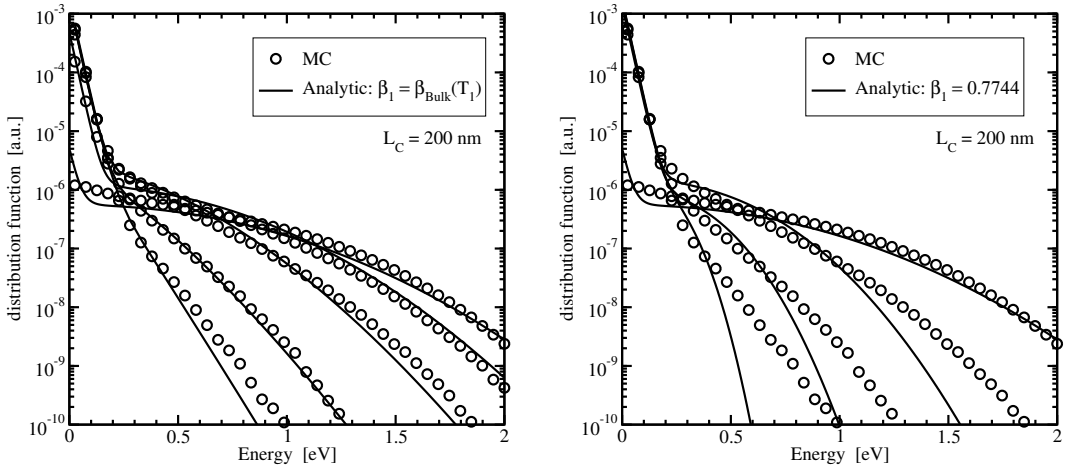

Figure 17: Comparison of analytical expressions for the distribution function with Monte Carlo results at different positions inside an n^+-n-n^+ test-structure with $L_\mathrm{C} = 200\ \mathrm{nm}$. Note the error in the tail when a constant value for β_1 is assumed (bottom figure).

less populated than in the bulk case. Therefore we detect the drain region when $\beta_n > \beta_\mathrm{Bulk}(T_n)$ is fulfilled. Inside the channel we assume $c = 0$ because no cold subpopulation exists. Inside the drain region, however, we need to determine c and T_2 to explicitly allow for two separate populations.

Thus for each grid-point the following nonlinear equation system is solved using Newton's method

$$\begin{pmatrix} T_n(\mathcal{E}_\mathrm{ref}, b, c) \\ \beta_n(\mathcal{E}_\mathrm{ref}, b, c) \\ \beta_1(\mathcal{E}_\mathrm{ref}, b, c) \end{pmatrix} = \begin{pmatrix} T_n^\mathrm{MC} \\ \beta_n^\mathrm{MC} \\ \beta_\mathrm{Bulk}(T_1(\mathcal{E}_\mathrm{ref}, b, c)) \end{pmatrix} \tag{7}$$

Note that T_n, β_n, and β_1 are analytic expressions derived from the moments of (5) and T_n^MC and β_n^MC were taken from Monte Carlo simulations. As stated above, in the channel $c = 0$ is assumed and the last row of (7) is dropped. Since there is no cold population $T_1 = T_n$ holds inside the channel. This also holds at the transition point which guarantees a continuous transition between the two regions. A comparison with distribution functions obtained by Monte Carlo simulations is shown in Fig. 17. Note that a constant value for $\beta_1 = \beta_h$ as used in [19] underestimates the tail of the distribution function and thus the associated impact ionization rate. Furthermore, an approach based on a constant β_1 works only for the high field case because otherwise β_n will never reach β_h and the model erroneously creates a cold population throughout the whole device. For intermediate bias conditions, a spurious cold population would be predicted in the larger part of channel.

A closed form macroscopic impact ionization rate is then obtained by integrating Keldysh's expression [20]

$$P_\mathrm{II}(\mathcal{E}) = P_0 \Big(\frac{\mathcal{E} - \mathcal{E}_\mathrm{th}}{\mathcal{E}_\mathrm{th}} \Big)^2 \tag{8}$$

with (3) and (4) as

$$G_{\mathrm{II},l} = \int_{\mathcal{E}_{\mathrm{th}}}^{\infty} \mathcal{E}^l \, P_{\mathrm{II}}(\mathcal{E}) \, f(\mathcal{E}) \, g(\mathcal{E}) \, \mathrm{d}\mathcal{E} \tag{9}$$

$$\approx \int_{\mathcal{E}_{\mathrm{th}}}^{\infty} \mathcal{E}^l \, P_{\mathrm{II}}(\mathcal{E}) \, f_1(\mathcal{E}) \, g_{\mathrm{c}}(\mathcal{E}) \, \mathrm{d}\mathcal{E} \tag{10}$$

$$= n \, P_0 \, \mathcal{E}_{\mathrm{ref}}^l \left(\frac{\mathcal{E}_{\mathrm{ref}}}{\mathcal{E}_{\mathrm{C}}} \right)^{\lambda - 1/2} \frac{\Gamma_{1,l} - 2 {z_{\mathrm{th}}}^{-1/b} \, \Gamma_{2,l} + {z_{\mathrm{th}}}^{-2/b} \, \Gamma_{3,l}}{\Gamma\left(\frac{2l+3}{2b}\right) + (\eta \, \mathcal{E}_{\mathrm{ref}})^{\zeta} \, \Gamma\left(\frac{2l+2\zeta+3}{2b}\right)} \tag{11}$$

where $\Gamma(a, z)$ is the incomplete Gamma function and

$$\Gamma_{j,l} = \Gamma\left(\frac{j + l + \lambda}{b}, z_{\mathrm{th}}\right) \tag{12}$$

$$z_{\mathrm{th}} = \left(\frac{\mathcal{E}_{\mathrm{th}}}{\mathcal{E}_{\mathrm{ref}}}\right)^b \tag{13}$$

$$g_{\mathrm{c}}(\mathcal{E}) = g_0 \sqrt{\mathcal{E}_{\mathrm{C}}} \left(\frac{\mathcal{E}}{\mathcal{E}_{\mathrm{C}}}\right)^{\lambda} \tag{14}$$

Equation (14) is an approximation of the high-energy region of the density of states ($\mathcal{E}_{\mathrm{C}} = 0.35\,\mathrm{eV}$ and $\lambda = 1.326$), based on Cassi and Riccò's [21] model, which has been introduced to simplify the final expression. Furthermore, it is assumed that only $f_1(\mathcal{E})$ contributes to the impact ionization rate. In (11) $l = 0, 1, 2$ and denotes the entries for the continuity, energy balance, and kurtosis balance equations, respectively.

The final expression (11) is equivalent to the expression given in [16], but the parameters $\mathcal{E}_{\mathrm{ref}}$ and b are now calculated in a different way. We use the same parameter values as in the Monte Carlo simulation ($P_0 = 4.18 \times 10^{12}\ \mathrm{s}^{-1}$, $\mathcal{E}_{\mathrm{th}} = 1.12\ \mathrm{eV}$), values which fit available experimental data for bulk. The accuracy for the bulk case is the same as in [16] as there is no cold population ($c = 0$). A comparison with Monte Carlo simulations for n^+-n-n^+ test-structures is shown in Fig. 18 where the improvement to the model of [16] is obvious. In addition, the results obtained by assuming a constant β_1 and the results obtained by a Maxwellian distribution function ($\mathcal{E}_{\mathrm{ref}} = \mathrm{k_B} T_n$, $b = 1$, and $c = 0$) are shown which underline the importance of an accurate distribution function model for impact ionization rate modeling.

4.3 Conclusion

The new macroscopic impact ionization model has been proven to deliver accurate results for the homogenous case [16] and inhomogeneous cases down to nanoscale devices. In particular, the tail of the impact ionization rate inside the drain area is correctly predicted which is not possible with models based on the average carrier energy only. The additional accuracy is provided by the kurtosis of the distribution function which can be obtained via a six moments transport model, that is, a model which is one order higher than conventional energy-transport models. It is important to point out that the same parameters as in the Monte Carlo simulation were used to evaluate the model, so no fitting has been performed. Although we have taken a rather simple expression for the microscopic scattering rate which is known to be not very accurate for energies close to the threshold energy, an extension to more accurate models is straightforward. As the model involves only state variables of the equation system it is well suited for the implementation into conventional numerical device simulators.

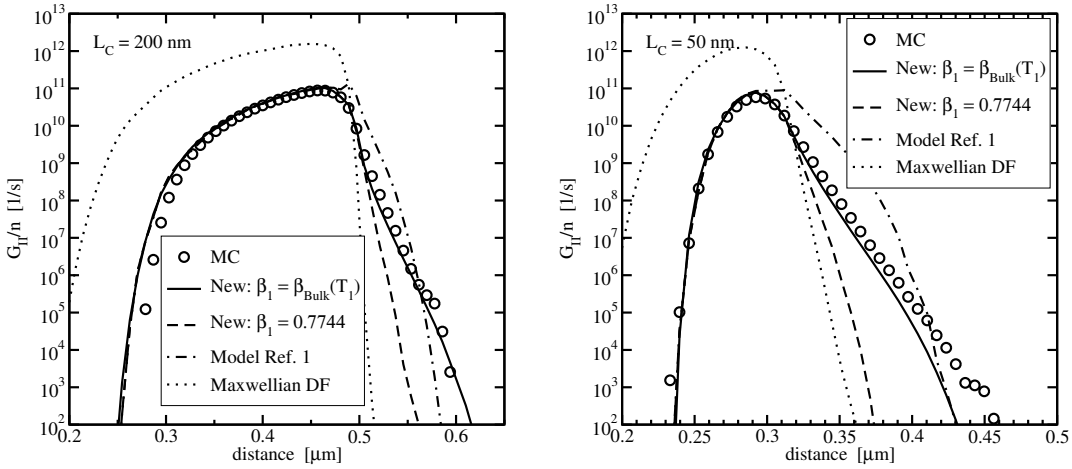

Figure 18: Comparison of analytically obtained impact ionization rates with Monte Carlo results for two n^+-n-n^+ test-structures. Note the improvement to the model of [16] and the large error when a Maxwellian distribution function is assumed ($E_{\max} = 300\ \mathrm{kV/cm}$).

Acknowledgement

This work has been supported by *Intel Corp.*, Santa Clara.

References

[1] W. Pyka, R. Martins, and S. Selberherr. Optimized Algorithms for Three-Dimensional Cellular Topography Simulation. *IEEE J.Technology Computer Aided Design*, 2000. `http://www.ieee.org/products/online/journal/tcad/accepted/Pyka-March00/`.

[2] E. Strasser and S. Selberherr. Algorithms and Models for Cellular Based Topography Simulation. *IEEE Trans.Computer-Aided Design*, 14(9):1104–1114, 1995.

[3] T. Binder and S. Selberherr. Object-Oriented Wafer-State Services. In R. V. Landeghem, editor, *14th European Simulation Multiconference*, pp 360–364, Ghent, Belgium, 2000. Society for Computer Simulation International.

[4] P. Lindstrom and G. Turk". Fast and Memory Efficient Polygonal Simplification. In *Proceedings of IEEE Visualization*, volume 98, pp 279–286, 1998.

[5] M. Gritsch, H. Kosina, T. Grasser, and S. Selberherr. Influence of Generation/Recombination Effects in Simulations of Partially Depleted SOI MOSFETs. *Solid-State Electron.*, 2001. (in print).

[6] M. Gritsch, H. Kosina, T. Grasser, and S.Selberherr. A Simulation Study of Partially Depleted SOI MOSFETs. In *Silicon-On-Insulator Technology and Devices X*, pp 181–186, Washington DC, USA, 2001.

[7] T. Simlinger, H. Brech, T. Grave, and S. Selberherr. Simulation of Submicron Double-Heterojunction High Electron Mobility Transistors with MINIMOS-NT. *IEEE Trans.Electron Devices*, 44(5):700–707, 1997.

[8] DESSIS-ISE Users Manual, Release. 6.

[9] J.L. Egley, B. Polsky, B. Min, E. Lyumkis, O. Penzin, and M. Foisy. SOI Related Simulation Challenges with Moment Based BTE Solvers. In *Simulation of Semiconductor Processes and Devices*, pp 241–244, Seattle, Washington, USA, 2000.

[10] D.J. Eaglesham, P.A. Stolk, H.J. Gossmann, and J.M. Poate. Implantation and Transient B Diffusion in Si: The Source of Interstitials. *Appl.Phys.Lett.*, 65(18):2305–2307, 1994.

[11] J.M. Poate, D.J. Eaglesham, G.H. Gilmer, J.-J. Gossmann, M. Jaraiz, C.S. Rafferty, and P.A. Stolk. Ion Implantation and Transient Enhanced Diffusion. In *Proc.Intl.Electron Devices Meeting*, pp 77–80, Washington, DC, USA, 1995.

[12] Avant! Corporation, TCAD Business Unit, Fremont, California, USA. *TSUPREM-4, Two-Dimensional Process Simulation Program, Version 2000.2 User's Manual*, 2000.

[13] C. Heitzinger and S. Selberherr. An Extensible TCAD Optimization Framework Combining Gradient Based and Genetic Optimizers. In *Proc. SPIE International Symposium on Microelectronics and Assembly: Design, Modeling, and Simulation in Microelectronics*, pp 279–289, Singapore, 2000.

[14] R. Strasser. *Rigorous TCAD Investigations on Semiconductor Fabrication Technology*. Dissertation, Technische Universität Wien, 1999. `http://www.iue.tuwien.ac.at/phd/strasser`.

[15] C.S. Rafferty, G.H. Gilmer, J. Jaraiz, D. Eaglesham, and H.-J. Gossmann. Simulation of Cluster Evaporation and Transient Enhanced Diffusion in Silicon. *Appl.Phys.Lett.*, 68(17):2395–2397, 1996.

[16] T. Grasser, H. Kosina, and S. Selberherr. Influence of the Distribution Function Shape and the Band Structure on Impact Ionization Modeling. *J.Appl.Phys.*, 90(12):6165–6171, 2001.

[17] T. Grasser, H. Kosina, M. Gritsch, and S. Selberherr. Using Six Moments of Boltzmann's Transport Equation for Device Simulation. *J.Appl.Phys.*, 90(5):2389–2396, 2001.

[18] E.O. Kane. Band Structure of Indium Antimonide. *J.Phys.Chem.Solids*, 1:249–261, 1957.

[19] K. Sonoda, S.T. Dunham, M. Yamaji, K. Taniguchi, and C. Hamaguchi. Impact Ionization Model Using Average Energy and Average Square Energy of Distribution Function. *Japanese Journal of Applied Physics*, 35(2B):818–825, 1996.

[20] L.V. Keldysh. Concerning the Theory of Impact Ionization in Semiconductors. *Sov.Phys.JETP*, 21:1135–1144, 1965.

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