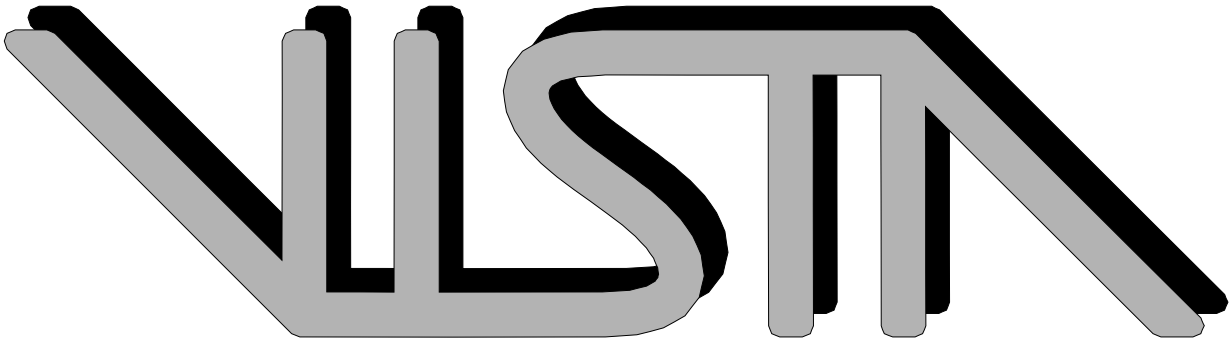


VISTA Status Report
June 2005

R. Entner, R. Heinzl, Ch. Hollauer,
A. Sheikholeslami, R. Wittmann, S. Selberherr

Institute for Microelectronics
Technical University Vienna
Gußhausstraße 27-29
A-1040 Wien, Austria

Contents

1 Impact of Multi-Trap Assisted Tunneling on Gate Leakage of CMOS Memory Devices

Dielectrics of state-of-the-art memory cells subject to repeated high field stress can have a high defect density. Thus, not only direct tunneling but also trap-assisted tunneling plays an important role. In this work a new approach for modeling gate leakage currents through highly degraded dielectrics is proposed. By rigorous simulation we show that multi-trap assisted tunneling becomes important for highly degrad dielectrics with thicknesses above approximately 4 nm, there it exceeds the single-trap assisted and direct tunneling components.

1.1 Introduction

While logic CMOS devices feature dielectric thicknesses below 1.2 nm, non-volatile memory cells rely on tunneling oxides as thick as 7 nm. In order to speed up the programming and erasing process strong electric fields are applied across the dielectric. Due to the repeated high-field stress, trap centers in the insulator are created, which lead to trap-assisted tunneling at low bias, forming stress-induced leakage current (SILC).

Modeling this gate leakage current for such devices is of paramount interest, because it determines the retention time. Thicker dielectrics subject to high field stress may have a high defect density. Thus not only direct tunneling but also trap-assisted tunneling (TAT) currents play an important role [1]. The trap-assisted current component has been found to stem from inelastic tunneling assisted by phonon emission [2].

For device simulation, trap-assisted tunneling is commonly modeled as a single-trap [3] or two-trap [4] process. Recently, multi-trap models considering hopping processes have been presented [5]. The single-trap model was found to accurately reproduce experimental data of slightly stressed dielectrics [6]. Recently, however, anomalous charge loss in floating-gate memory cells after program/erase stress cycles has been observed [7]. Due to the high defect density in those cells it is reasonable to assume that more than one trap is involved in the tunneling process. For correct modeling of such highly degraded devices a new approach is presented which rigorously computes TAT current assisted by multiple traps. In contrast to the model presented in [5], where conduction across discrete paths is assumed, hopping processes between all traps are taken into account. The space charge of occupied traps is accounted for in the Poisson equation to estimate the resulting V_t shift.

1.2 The Model

For the simulation of trap-assisted tunneling currents the current density across the insulator is modeled as the sum of the capture and emission rates R_i in each trap times the trap cross section Δx_i,

$$J = q \sum_i R_i \Delta x_i. \tag{1}$$

The energetic position of the trap with respect to the conduction band edge $\mathcal{E}_\mathrm{T}$ determines the trap cross section [8]

$$\Delta x_i = \frac{\hbar}{\sqrt{2 m_\mathrm{diel} \mathcal{E}_\mathrm{T}}} \left(\frac{4\pi}{3}\right)^{1/3}, \tag{2}$$

where m_diel denotes the electron mass in the dielectric, which is used as a fitting parameter.

The single-TAT and the multi-TAT models differ in the way R_i is calculated. When only single-trap processes are considered (see Figure 1) the rates are determined by [9]

$$R_{\mathrm{c}_i} = \tau_{\mathrm{c}_i}^{-1} N_{\mathrm{t}_i} (1 - f_{\mathrm{t}_i}) \;,\; R_{\mathrm{e}_i} = \tau_{\mathrm{e}_i}^{-1} N_{\mathrm{t}_i} f_{\mathrm{t}_i}. \tag{3}$$

Here, R_{c_i} and R_{e_i} are the capture and emission rates of the considered trap, respectively, and N_{t_i} denotes the trap concentration. In the stationary case the capture and emission rates must be equal, hence $R_{\mathrm{c}_i} = R_{\mathrm{e}_i} = R_i$. The trap occupancy f_{t_i} can be directly calculated as $f_{\mathrm{t}_i} = \tau_{\mathrm{c}_i}^{-1} / (\tau_{\mathrm{c}_i}^{-1} + \tau_{\mathrm{e}_i}^{-1})$ where the inverse capture and emission times can be evaluated as [3, 9]

$$\tau_{c_i}^{-1} = \int_{\mathcal{E}_0}^{\infty} g_C(\mathcal{E}) c_n(\mathcal{E}) T_C(\mathcal{E}) f_C(\mathcal{E}) \, d\mathcal{E}, \quad (4)$$

$$\tau_{e_i}^{-1} = \int_{\mathcal{E}_0}^{\infty} g_A(\mathcal{E}) e_n(\mathcal{E}) T_A(\mathcal{E}) (1 - f_A(\mathcal{E})) \, d\mathcal{E}. \quad (5)$$

In these expressions, $g_C(\mathcal{E})$ and $g_A(\mathcal{E})$ denote the density of states in the cathode and anode, respectively, and the symbols c_n and e_n are computed as

$$c_n(\mathcal{E}) = c_0 \sum_m L_m \delta(\mathcal{E} - \mathcal{E}_m), \quad (6)$$

$$e_n(\mathcal{E}) = c_0 \exp\left(-\frac{\mathcal{E} - \mathcal{E}_T}{k_B T_L}\right) \sum_m L_m \delta(\mathcal{E} - \mathcal{E}_m) \quad (7)$$

with

$$c_0 = (4\pi)^2 \Delta x_i^2 (\hbar\Theta)^3 / (\hbar^2 \mathcal{E}_{g,SiO2}), \quad (8)$$

$$(\hbar\Theta) = (q^2 \hbar^2 F^2 / (2\, m_{diel}))^{1/3}. \quad (9)$$

The summation index m denotes the discrete phonon emissions, $\mathcal{E}_m$ is the phonon energy, and L_m is the multiphonon transition probability [9]. The symbols f_c and f_a are the Fermi distributions, T_c and T_a the transmission coefficients from the cathode and the anode, F the electric field in the dielectric, and $\mathcal{E}_{g,SiO2}$ the band gap of SiO_2. The transmission coefficients were evaluated by a numerical WKB method, which yields reasonable accuracy for single-layer dielectrics. This model has been used in a more or less similar form by various authors [1, 2, 3].

Recently, however, anomalous charge loss in memory cells has been observed and was explained by conduction through a second trap [4]. The single-trap model can be extended for this case, and the rate equations become (see Figure 2)

$$\overbrace{\tau_{c_{0,1}}^{-1} N_{t_1}(1 - f_{t_1})}^{R_{c_1}} - \overbrace{(\tau_{e_{1,2}}^{-1} N_{t_1} f_{t_1}(1 - f_{t_2}) + \tau_{e_{1,3}}^{-1} N_{t_1} f_{t_1})}^{R_{e_1}} = 0,$$

$$\underbrace{\tau_{c_{0,2}}^{-1} N_{t_2}(1 - f_{t_2}) + \tau_{c_{1,2}}^{-1} N_{t_2} f_{t_1}(1 - f_{t_2})}_{R_{c_2}} - \underbrace{\tau_{e_{2,3}}^{-1} N_{t_2} f_{t_2}}_{R_{e_2}} = 0,$$

where instantaneous transitions between occupied and free traps are assumed. For thicker dielectrics it is quite reasonable to assume that an arbitrary number of traps assists in the conduction process. We therefore extend the model to n traps where the

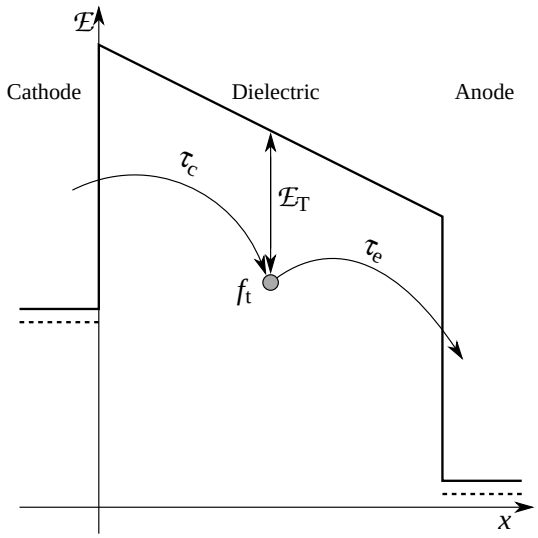


Figure 1: Single-trap assisted tunneling process. The tunneling rate R_i of a specific trap is determined by the capture time from the cathode and the emission time to the anode.

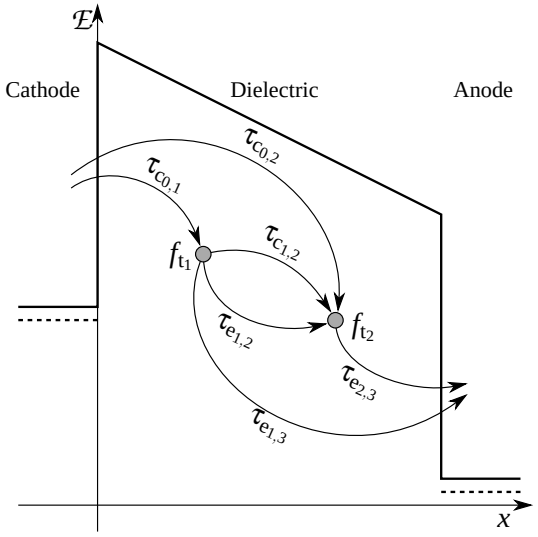


Figure 2: Multi-trap assisted tunneling process. The tunneling rate R_i of a specific trap is determined by all capture and emission times to and from the trap.

capture and emission rates are evaluated as

$$R_{c_k} = \sum_{i=0}^{k-1} \tau_{c_{i,k}}^{-1} N_{t_k} f_{t_i}(1 - f_{t_k}), \quad (10)$$

$$R_{e_k} = \sum_{i=k+1}^{n+1} \tau_{e_{k,i}}^{-1} N_{t_k} f_{t_k}(1 - f_{t_i}). \quad (11)$$

The values for f_{t_0} and f_{t_n}, which are the trap occupation probabilities at the cathode and the anode, are set to 1 and 0 respectively. This way the cathode acts as electron source and the anode as electron sink. The values for the other trap occupation probabilities have to be evaluated from the equation system. This is performed within MINIMOS-NT using the Newton method. Typical dimensions of the equation system to be solved are, depending on the dielectric thickness and trap energy, up to 15×15. The computational effort remains negligible compared to the total device simulation time. From either the capture or emission rates the multi-trap assisted tunneling current density J is obtained.

1.3 Application

The implementation of these models into the device- and circuit-simulator MINIMOS-NT [10] allows the two- and three-dimensional study of single- and multi-trap assisted tunneling. Figure 3 shows measured SILC [6] after different stress times for a MOS capacitor and the representative simulation results using a single-TAT simulation. It can be clearly seen that for slightly degraded dielectrics the single-TAT model yields excellent agreement with the measured data. Here the trap concentration is used as fitting parameter. For highly degraded devices it has been shown [4], however, that the SILC cannot be explained by conduction over solitary traps. It must be assumed that the large SILC is due to defect interaction of nearby traps.

Figure 4 shows a comparison of the three different tunneling mechanisms, namely direct tunneling, single-TAT, and multi-TAT. The models have been applied to a set of MOS transistors with gate dielectric thicknesses ranging from 1.5 nm up to 9 nm. The gate is biased at 1 V, source and drain are kept at 0 V. For both, the single-TAT and the multi-TAT simulations, the trap energy is set to 2.8 eV below the dielectric conduction band with a constant trap density of $9 \times 10^{17}\,\mathrm{cm}^{-3}$ across the oxide. It can be clearly seen that in the multi-trap simulation the tunneling current is several orders of magnitude higher than in the single-trap simulation. This is due to the fact that the multi-

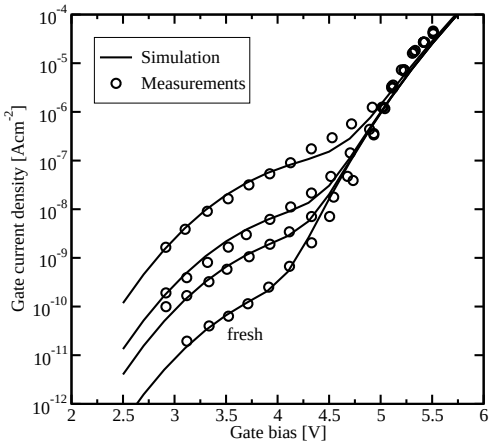


Figure 3: Single-TAT simulations of a MOS capacitor with a 5.5 nm dielectric and N_t=$9 \times 10^{17}\,\mathrm{cm}^{-3} \ldots 3 \times 10^{15}\,\mathrm{cm}^{-3}$ (top to bottom).

TAT current includes the single-TAT component as a limiting case. The multi-TAT model considers the capture and emission processes from the cathode and to the anode, respectively, but also the capture and emission processes involving all other trap centers. This fact leads to the comparably high multi-TAT component in devices with thicker oxides. It has to be considered, though, that this high current is mainly due to the assumption of uniformly distributed trap concentrations across the oxide. The direct tunneling component loses importance for thicker dielectrics but dominates for thin dielectrics as found in logic CMOS devices. For miniaturized devices with thicker oxides and higher trap densities multi-TAT processes become increasingly important.

Figure 5 depicts the trap occupancy within the oxide. A MOS transistor with 1 V gate bias was simulated. It can be seen that the trap occupancy f_t is remarkably lower in the multi-TAT case. The reason is the higher probability for electrons to tunnel to one of the neighbor traps compared to tunneling to the anode as it is the only possibility in the single-TAT model.

The implementation of this model into the device simulator MINIMOS-NT allows to simulate the effect on the threshold voltage of memory devices. The space charge density in the dielectric is calculated as $\rho(x) = q f_t(x) N_t(x)$. Figure 6 outlines

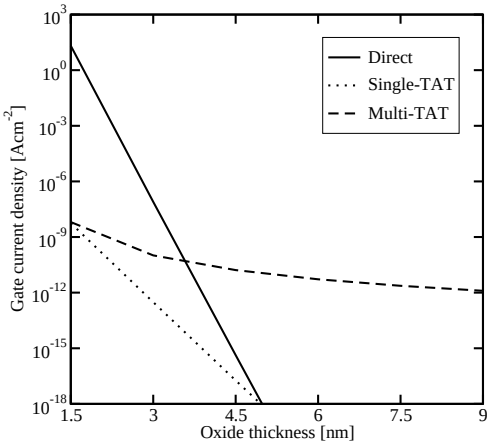


Figure 4: SILC simulations for a set of MOS devices at 1 V gate bias. The oxide has a constant trap concentration of $9 \times 10^{17}\,\mathrm{cm}^{-3}$.

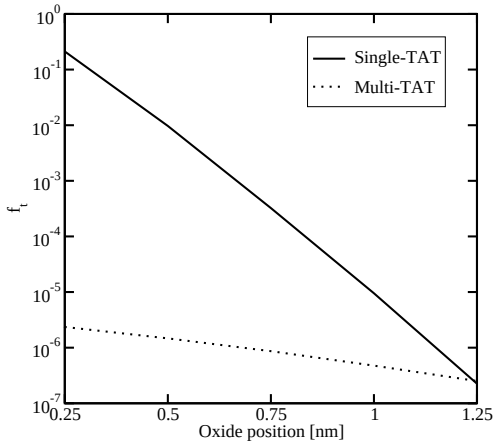


Figure 5: The trap occupancy f_t in the oxide of a 1.5 nm MOS transistor at 1 V gate bias.

the threshold voltage V_t for different oxide thicknesses. The direct tunneling model, applying the commonly used Tsu-Esaki approach, does not account for the filling of traps in the oxide. Therefore the threshold voltage is not shifted compared to the simulation without a tunneling model. The new multi-TAT model predicts an increase in V_t. This higher threshold voltage is due to the tunneling current in the oxide and the filled and therefore negatively charged traps.

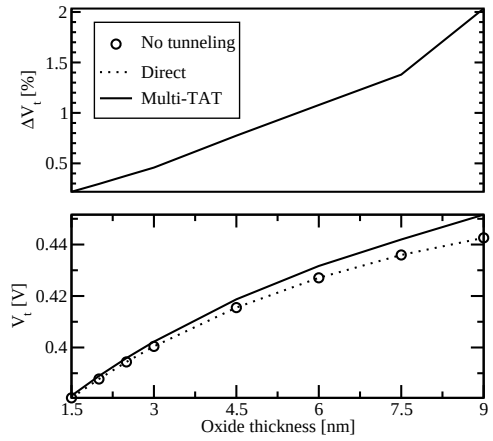


Figure 6: Comparison of the threshold voltage V_t of MOSFET structures with different oxide thicknesses.

1.4 Conclusion

We presented a new trap-assisted tunneling model which takes the interaction of several traps for the creation of conducting paths into account. The model is applied to devices with varying oxide thicknesses. Comparing single-TAT and direct tunneling reveals that for highly degraded devices with oxide thicknesses between 3 nm and 8 nm the inclusion of a multi-TAT model is crucial. With the implementation of the multi-TAT model in the multi-purpose device- and circuit simulator MINIMOS-NT arbitrary device geometries can be evaluated.

2 Three-Dimensional Simulation of Thermal Oxidation and the Influence of Stress

The thermal oxidation process of three-dimensional structures is analyzed with our oxidation model. This comprehensive model takes into account that the diffusion of oxidants, the chemical reaction, and the volume increase occur simultaneously in a so-called reactive layer which has a spatial finite width, in contrast to the sharp interface between silicon and dioxide in the conventional formulation. Our oxidation model also includes the coupled stress dependence of the oxidation process because the influence of stress is shown to be considerable. Only the simulation of stress dependent oxidation leads to results which agree with the real physical behavior.

2.1 Introduction

Thermal oxidation of silicon is one of the most important steps in the fabrication of highly integrated electronic circuits, being mainly used for efficient isolation of adjacent devices from each other. If a surface of a silicon body has contact with an oxidizing atmosphere, the chemical reaction of oxidants with silicon (Si) forms silicon dioxide (SiO_2). Depending on the oxidizing ambient the oxidants can arise from steam (dry oxidation) or water vapour (wet oxidation). Wet oxidation has a significantly higher oxidation rate than the dry one and so wet oxidation is mainly applied for fast thick film oxidation in contrast to dry oxidation which is more suitable for thin film oxidation.

If a SiO_2 - domain already exists the oxidants diffuse through the SiO_2 - domain to the Si - SiO_2 - interface [11] where the chemical reaction of the oxidants with silicon takes place. The parts of silicon which should not be oxidized are masked by a layer of silicon nitride (Si_3Ni_4), because Si_3Ni_4 prevents the oxidant diffusion on the subjacent SiO_2 - layer. During the oxidation process the chemical reaction consumes Si and the newly formed SiO_2 has more than twice the volume of the original Si. This significant volume increase leads to large displacements in the materials.

If the additional volume is prevented from expanding as desired, e.g. by the Si_3Ni_4 - mask, mechanical stresses arise in the materials. Since there is a strong stress dependence of the oxidant diffusion and the chemical reaction, the oxidation process itself is considerably influenced by stress.

So thermal oxidation is a complex process where the three subprocesses oxidant diffusion, chemical reaction, and volume increase occur simultaneously. From a mathematical point of view the oxidation process can be described by a coupled system of partial differential equations, one for the diffusion of oxidants through SiO_2, the second for the conversion of Si into SiO_2 at the interface, and a third for the mechanical problem of the complete oxidized structure. The whole mathematical formulation is numerically solved by applying the finite element method.

The oxidation rate mainly depends on the oxidizing ambient, especially on the temperature, the pressure, and the oxidant species. For practical use of an oxidation model it is important to carefully control the ambient parameters of the oxidation process. Therefore in our model these parameters, e.g the temperature profile, are adjustable easily.

2.2 The Model

In our oxidation model [12] we use a normalized silicon concentration $\eta(\vec{x},t)$ [13] so that the value of η is 1 in pure Si and 0 in pure SiO_2. Advantageously our model takes into account that the diffusion of oxidants, the chemical reaction and the volume increase occur simultaneously in a so-called reaction layer. In contrast to the sharp interface between Si and SiO_2 like in the standard model [14], this reaction layer has a spatial finite width (see Figure 9) where the value of η lies between 0 and 1.

Most of the other oxidation models [15][16] describe the SiO_2 - growth more or less by a moving bound-

ary problem, because these models are all based on the one-dimensional standard model [14]. Unfortunately the handling of moving boundary problems becomes very complicated for complex three-dimensional structures [17] and so sometimes such models are restricted to two dimensions or have a lack of reliability and stability for complex three-dimensional structures.

With regard to these aspects our oxidation model is advantageous, because with the normalized silicon concentration and the reaction layer it exhibits always the same reliability independent of the geometry of the three-dimensional structure. Furthermore all other oxidation models need a special fitting for thin film oxidation as described in [18]. So another advantage of or model is that thin film or dry oxidation [19][20] is also properly treated by our model without any modification.

2.2.1 Oxidant Diffusion

The diffusion of oxidants is described by

$$D(p)\,\Delta C(\vec{x},t) = k(\eta,p)\,C(\vec{x},t), \qquad (12)$$

where Δ is the Laplace operator, $C(\vec{x},t)$ is the oxidant concentration, and $D(p)$ is the stress dependent diffusion coefficient:

$$D(p) = D_0 \exp\!\left(-\tfrac{pV_D}{k_B T}\right). \qquad (13)$$

Here D_0 is the low stress diffusion coefficient [21][22], p is the hydrostatic pressure in the respective material, V_D is the activation volume, k_B is the Boltzmann's constant, and T is the temperature in Kelvin.
$k(\eta,p)$ is the stress dependent strength of a spatial sink and not just a reaction coefficient at a sharp interface:

$$k(\eta,p) = \eta(\vec{x},t)\,k_{max} \exp\!\left(-\tfrac{pV_k}{k_B T}\right). \qquad (14)$$

As given in (14) we define that $k(\eta,p)$ is linearly proportional to $\eta(\vec{x},t)$. Furthermore (13) and (14) are only valid for a pressure $p \geq 0$.

2.2.2 Dynamics of η

The dynamics of η is described by

$$\frac{\partial \eta(\vec{x},t)}{\partial t} = -\frac{1}{\lambda} k(\eta,p)\,C(\vec{x},t)/N_1, \qquad (15)$$

where λ is the volume expansion factor (= 2.25) for the reaction from Si to SiO_2, and N_1 is the number of oxidant molecules incorporated into one unit volume of SiO_2.

2.2.3 Mechanics

The chemical reaction of silicon and oxygen causes a significant volume increase. The normalized additional volume after time Δt is

$$V_{rel}^{add} = \frac{\lambda - 1}{\lambda} \Delta t\; k(\eta,p) C(\vec{x},t)/N_1. \qquad (16)$$

The mechanical problem is described by the equilibrium relations

$$\begin{aligned} \frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{xy}}{\partial y} + \frac{\partial \sigma_{xz}}{\partial z} &= 0 \\ \frac{\partial \sigma_{yx}}{\partial x} + \frac{\partial \sigma_{yy}}{\partial y} + \frac{\partial \sigma_{yz}}{\partial z} &= 0 \\ \frac{\partial \sigma_{zx}}{\partial x} + \frac{\partial \sigma_{zy}}{\partial y} + \frac{\partial \sigma_{zz}}{\partial z} &= 0 \end{aligned} \qquad (17)$$

where the stress tensor $\tilde{\sigma}$ is given by

$$\tilde{\sigma} = \mathbf{D}(\tilde{\varepsilon} - \tilde{\varepsilon_0}) + \tilde{\sigma_0}. \qquad (18)$$

Here **D** is the so-called material matrix. Furthermore, $\tilde{\varepsilon}$ is the strain tensor, $\tilde{\varepsilon_0}$ is the residual strain tensor, and $\tilde{\sigma_0}$ is the residual stress tensor.

The material matrix **D** can be splitted in a dilatation and a deviatoric part [23]

$$\mathbf{D} = \left(H \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} + G_{eff} \begin{bmatrix} +\frac{4}{3} & -\frac{2}{3} & -\frac{2}{3} & 0 & 0 & 0 \\ -\frac{2}{3} & +\frac{4}{3} & -\frac{2}{3} & 0 & 0 & 0 \\ -\frac{2}{3} & -\frac{2}{3} & +\frac{4}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \right) \qquad (19)$$

with the bulk modulus H and the effective shear modulus G_{eff}. The bulk modulus is

$$H = \frac{E}{3(1-2\nu)}, \tag{20}$$

where E is the Young modulus and ν is the Poisson ratio.

In the elastic case G_{eff} is the same as the standard shear modulus

$$G_{eff} = G = \frac{E}{2(1+\nu)}. \tag{21}$$

The materials SiO_2 and Si_3Ni_4 have a viscoelastic behavior [24]. The Maxwellian formulation of viscoelasticity is the most suitable one for these materials [25]. In Maxwell's model the dilatation part is assumed purely elastic while the deviatoric part is modelled by a Maxwell element.

It can be assumed that for a short time period ΔT the strain velocity can be kept constant ($\dot{\varepsilon} = \frac{\varepsilon}{\Delta T}$) [26]. So in the viscoelastic case G_{eff} can be written in the form

$$G_{eff} = G\frac{\tau}{\Delta T}\left(1 - \exp\left(-\frac{\Delta T}{\tau}\right)\right), \tag{22}$$

where τ is the Maxwellian relaxation time. This relationship shows that Maxwell viscoelasticity can be expressed by an effective rigidity G_{eff} in the deviatoric part of (19). So the material matrix **D** depends in the elastic case only on Young's modulus E and the Poisson ratio ν, and in the viscoelastic case additionally on the Maxwellian relaxation time τ.

The components $\varepsilon_{0,ii}$ (i stands for x, y or z) of the residual strain tensor $\tilde{\varepsilon_0}$ are linear proportional to the normalized additional volume from

$$\varepsilon_{0,ii} = \tfrac{1}{3}V_{rel}^{add}. \tag{23}$$

After discretization of the continuum, we obtain a linear equation system for the mechanical problem

$$\mathbf{K}\vec{d} = \vec{f} \quad \text{with} \quad \mathbf{K} = \int_{\mathcal{V}} \mathbf{B^T D B} dV, \tag{24}$$

where **K** is the so-called stiffness matrix, $\vec{d}$ is the displacement vector, and $\vec{f}$ is the force vector.

$$\mathbf{B} = [\mathbf{B_i}, \mathbf{B_j}, \mathbf{B_m}, \mathbf{B_p}] \tag{25}$$

is a discretized partial derivative matrix [27] where i, j, m and p are the four nodes on a single tetrahedral element. The submatrix $\mathbf{B_i}$ for the node i is

$$\mathbf{B_i} = \begin{bmatrix} \frac{\partial N_i}{\partial x} & 0 & 0 \\ 0 & \frac{\partial N_i}{\partial y} & 0 \\ 0 & 0 & \frac{\partial N_i}{\partial z} \\ \frac{\partial N_i}{\partial y} & \frac{\partial N_i}{\partial x} & 0 \\ 0 & \frac{\partial N_i}{\partial z} & \frac{\partial N_i}{\partial y} \\ \frac{\partial N_i}{\partial z} & 0 & \frac{\partial N_i}{\partial x} \end{bmatrix} = \begin{bmatrix} b_i & 0 & 0 \\ 0 & c_i & 0 \\ 0 & 0 & d_i \\ c_i & b_i & 0 \\ 0 & d_i & c_i \\ d_i & 0 & b_i \end{bmatrix}, \tag{26}$$

with the linear form function $N_i(\vec{x})$ defined as

$$N_i(\vec{x}) = a_i + b_i x + c_i y + d_i z, \tag{27}$$

in which a_i, b_i, c_i and d_i are constant geometrical coefficients for the finite element. For example b_i is

$$b_i = -\det \begin{vmatrix} 1, & y_j, & z_j \\ 1, & y_m, & z_m \\ 1, & y_p, & z_p \end{vmatrix}. \tag{28}$$

The force vector on a finite element depends on the residual strain tensor and thus also on the volume increase

$$\vec{f} = \int_{\mathcal{V}} \mathbf{B^T D}\tilde{\varepsilon}_0 \, dV. \tag{29}$$

After solving the linear equation system (24) we obtain the displacement vector. Since the strain is the first derivative with respect to displacement

$$\tilde{\varepsilon} = \mathbf{B}\vec{d}, \tag{30}$$

the stress can be determined with equation (18).

With the stress tensor the pressure for (13) and (14) can be determined by using the formula

$$p = -\frac{\text{Trace}(\tilde{\sigma})}{3} = -\frac{\sigma_{xx} + \sigma_{yy} + \sigma_{zz}}{3}. \tag{31}$$

2.3 Simulation Procedure

For the simulation procedure we perform a finite element discretization by splitting the three-dimensional structure into tetrahedral elements. The size of the tetrahedrons and, as a result of that, the number of finite elements is controlled adaptively by a meshing module.

In the next step we set the initial values for the oxidant concentration C and the normalized silicon concentration η on the grid nodes. For example η must be 1 in pure Si.

As shown in Figure 7, we iterate over all finite elements and build the local equation system for each element for every actual discrete time. The local equation system describes the oxidation process numerically for one element. In order to describe the oxidation process on the complete simulation domain we need a global coupled equation system. The components of the global equation system are assembled from the local equation system by using the superposition principle.

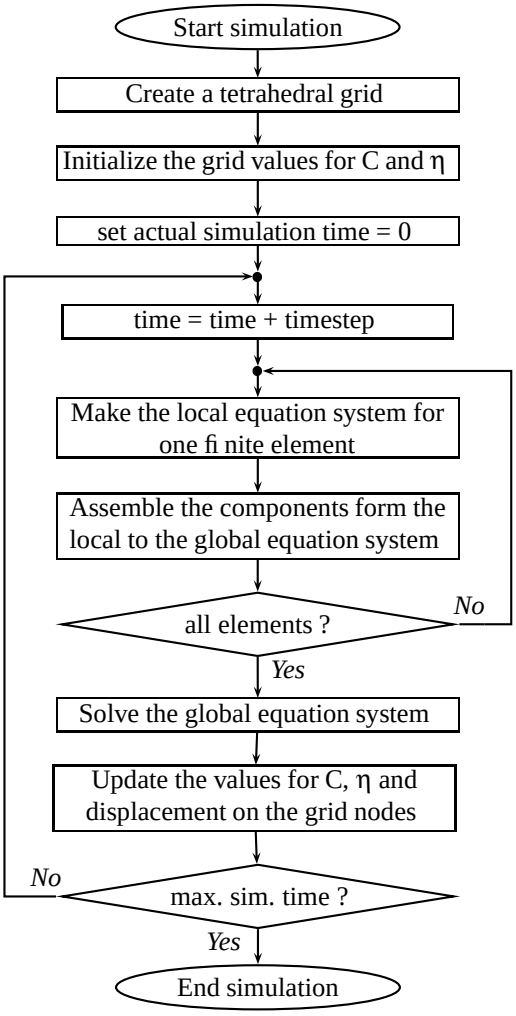


Figure 7: Simulation procedure

After the iteration over all elements is finished, the global assembled equation system is also completed. Now the global non-linear equation system can be solved and we obtain the results for C, η, and the displacement values for the actual time step. With these results we update the values for C, η, and the displacements on the grid nodes for the actual time step. such that these values are always keeping pace with the actual simulation time. The displacement vector enables the calculation of the strain tensor (30) as well as the stress tensor (18).

When the above described procedure is finished, we increase the actual simulation time and start with the assembling loop again. The same assembling and solving procedure is repeated for each time step until the desired end of the simulation.

2.4 Representative Example

2.4.1 First Example

We apply our oxidation model to the three-dimensional structure with (1.2×0.3) μm floor space, as displayed on Figure 8. In this example the upper layer is a 0.15 μm thick Si_3Ni_4 - mask which prevents the oxidant diffusion on the subjacent Si - layer. Here the bottom surface is fixed and on the upper surface a free mechanical boundary condition is applied. The result of the oxidation process of the whole structure after a time t_1 is shown in Figure 9.

For a more physical interpretation of the simulation results with a sharp interface between Si and SiO_2 the two regions can be extracted from the η - distribution by determining that $\eta \leq 0.5$ is SiO_2 and $\eta > 0.5$ is Si as shown in Figure 10. For an optimal comparison of the geometry before and after oxidation as well as the influence of stress, Figure 8–13 have the same perspectives and the same proportions.

2.4.2 Stress Dependence

In order to demonstrate the importance of the stress dependence we compare the results with and without the impact of stress. Since the oxidant diffusion and the chemical reaction are exponentially reduced with the hydrostatic

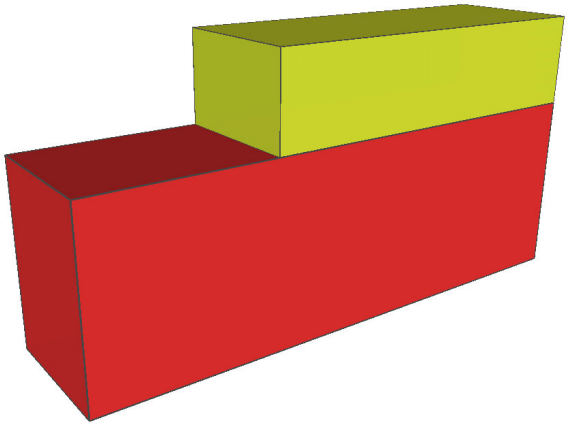


Figure 8: Initial structure of the Si - Si_3Ni_4 - body before thermal oxidation

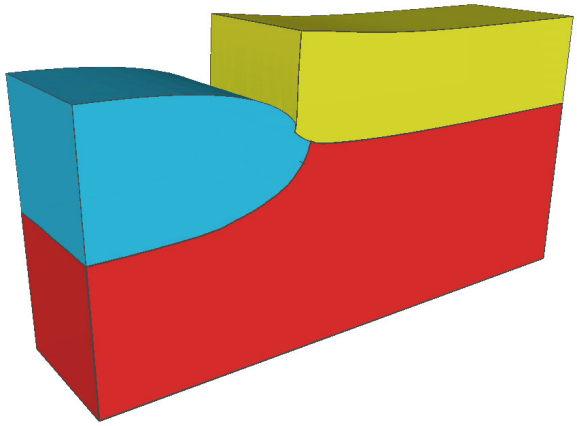


Figure 10: SiO_2 - region (sharp interface) with stress dependent oxidation at time t_1.

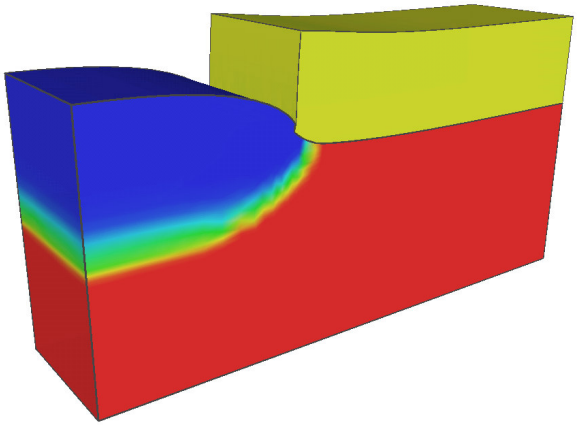


Figure 9: η - distribution and reaction layer after thermal oxidation at time t_1

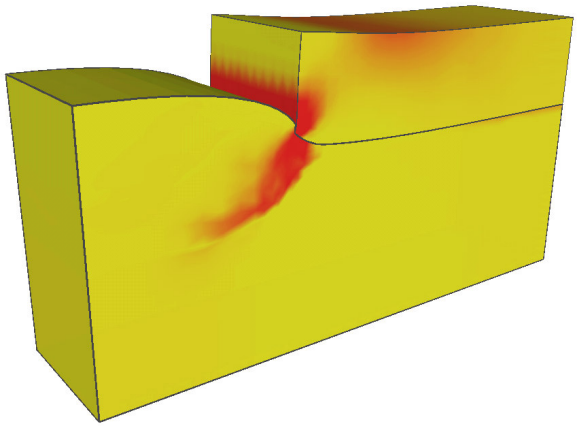


Figure 11: Pressure distribution with stress dependent oxidation at time t_1.

pressure in the material, the oxidation process itself is highly stress dependent.

As shown in Figure 11 the highest pressure in SiO_2 is under the edge of the Si_3Ni_4 - mask, because in this area the stiffness of the mask prevents the desired volume expansion of the newly formed SiO_2. Due to the mentioned stress dependence the oxidation rate in these areas is considerably reduced (see Figure 10).

If the stress dependence is not included in the simulation of the oxidation process, the simulation results do not agree with the real physical behavior, because the oxidant diffusion and the chemical reaction also occur under the Si_3Ni_4 - mask without restriction. Because of this phenomenon the SiO_2 - region at the same oxidation conditions is much more expanded than with the stress dependence as shown in Figure 12. In addition, the larger forces under the Si_3Ni_4 - mask, which result from the larger pressure domain (see Figure 13) in this area, cause larger displacements of the mask.

2.4.3 Second Example

Another more complex three-dimensional structure ((0.8×0.8) μm floor space), with a 0.15 μm thick L - shaped Si_3Ni_4 - mask is stress dependent oxidized with our oxidation model, as shown in Figure 14. Due to the L - shaped mask here the effect of the three-dimensional oxidation process is pronounced, because the shape of the SiO_2 - region and the deformations are not continuous in any direction.

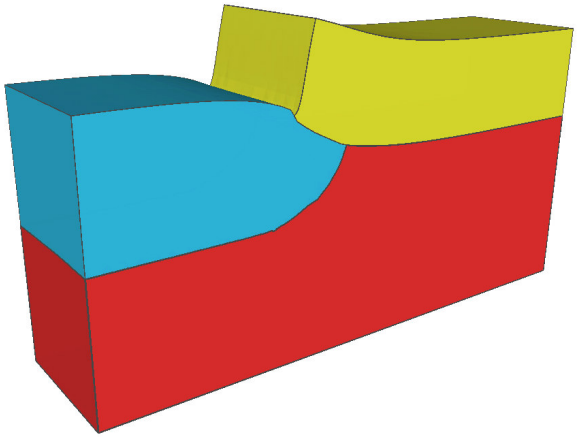


Figure 12: SiO_2 - region (sharp interface) without stress dependent oxidation at time t_1.

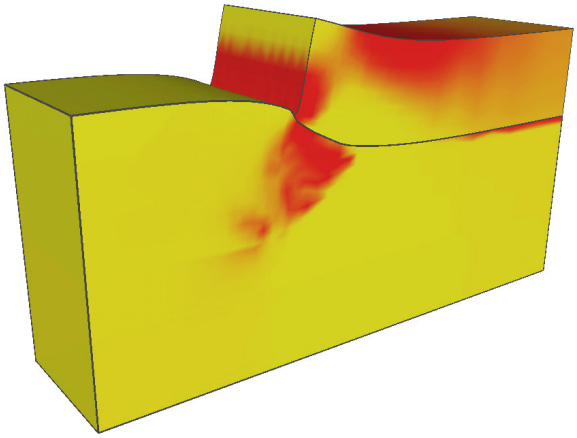


Figure 13: Pressure distribution without stress dependent oxidation at time t_1.

Figure 15 shows that the highest pressure in SiO_2 is under the edge of the Si_3Ni_4 - mask again, which slows down the oxidation process in these areas. The stiffness of the Si_3Ni_4 - mask is approximately six times larger than the stiffness of SiO_2 and so the displacements in SiO_2 are also much more larger than in the Si_3Ni_4 - mask which leads to the well-known bird's beak effect.

2.5 Summary and Conclusion

An oxidation model which takes the real physical behavior of the whole oxidation process under full control of the ambient parameters into account, has been presented. Our model is based on a normalized silicon concentration and a reaction layer with a spatial finite width in contrast to the moving boundary concept and a sharp Si - SiO_2 - interface in the conventional formulation.

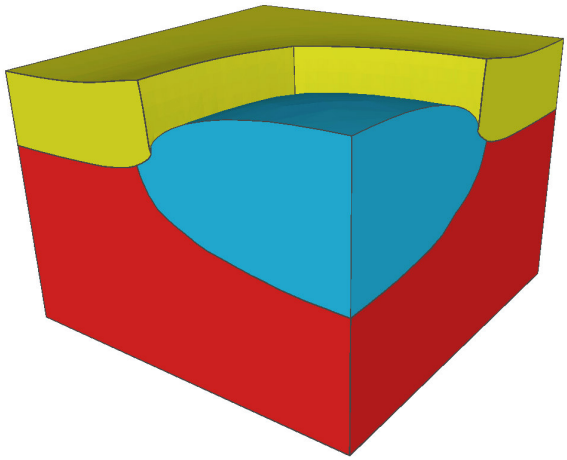


Figure 14: SiO_2 - region with stress dependent oxidation at time t_1.

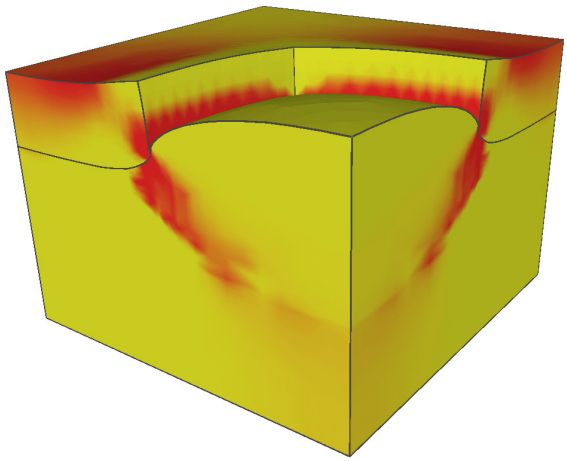


Figure 15: Pressure distribution with stress dependent oxidation at time t_1.

The reaction layer takes into account that the diffusion of oxidants, the chemical reaction and the volume increase occur simultaneously. In contrast to the moving boundary concept, our normalized silicon concentration concept also works on complex three-dimensional structures without restriction. For a physical interpretation with a sharp interface between Si and SiO_2 the two regions can be extracted from the normalized silicon distribution by determining that a value equal or less 0.5 is SiO_2 and a value larger than 0.5 is Si. Furthermore thick film as well as thin film oxidation are properly treated without a need of modification.

For the mechanical behavior of the materials an elastic or viscoelastic model can be applied. In the viscoelastic model we use use a Maxwellian formulation. It was shown that for short time periods the Maxwell viscoelasticity can be expressed by an effective shear modulus which depends on the Maxwellian relaxation time, in the deviatoric part of the so-called material matrix. The whole mathematical formulation of the oxidation process, which is described by a coupled system of partial differential equations, is solved by applying the finite element method.

The model was verified by the oxidation of two different three-dimensional structures. Beside the presentation of the simulation results the important influence of stress on thermal oxidation was investigated. It was shown that the highest hydrostatic pressure in SiO_2 is under the edge of the Si_3Ni_4 - mask and that the results agree only for stress dependent oxidation with the real physical behavior. As a result of the strong stress dependence of the oxidant diffusion and the chemical reaction the oxidation rate in this area is considerably reduced. If the stress is not taken into account the oxidation process also occurs under the Si_3Ni_4 - mask without restriction and so the SiO_2 - domain and the displacements are too large in relation to the real physical process.

3 A Novel Technique for 3D Mesh Adaptation with an A Posteriori Error Estimator

We present a novel error estimation driven three-dimensional unstructured mesh adaptation technique based on a posteriori error estimation techniques with upper and lower error bounds. In contrast to other work [28] we present this approach in three dimensions using unstructured meshing techniques to potentiate an automatically adaptation of three-dimensional unstructured meshes without any user interaction. The motivation for this approach, the applicability and usability is presented with real-world examples.

3.1 Introduction

Most TCAD (Technology Computer Aided Design) problems can be formulated with partial differential equations and solved by numerical methods, usually finite difference, finite element and finite volume methods. They are used to model disparate phenomena such as dopant diffusion, mechanical deformation, heat transfer, fluid flow, electromagnetic wave propagation, and quantum effects. An essential step in these methods is to find a proper tessellation of a continuous domain with discrete elements, in our case tetrahedra.

This transition from the continous domain to a discretized domain will inherently produce errors in the computed results, no matter how sophisticated or how appropriate a mathematical model is. This approximation error can be enormous, and can completely invalidate numerical predictions if we have no estimated or quantitative measurement of these errors.

The general subject is referred to as *a posteriori error estimation*. It is an essential step to observe and bound the approximation error and to have a mesh adaptation strategy to guarantee the accuracy of the solution within a given range. In contrast to two dimensions where mesh generation and adaptation techniques are mostly based on hand crafted meshes or grids, it is almost impossible to design grids or meshes in three dimensions. On that account it is very important to generate and adapt meshes in three-dimensions automatically.

3.2 Mesh Generation and Adaptation

The first step in solving equations numerically is the discretization of the underlying computational domain. A widely used approach has been to divide the domain into a structured assembly of quadrilateral cells, with the topological information being apparent from the fact that each interior vertex is surrounded by exactly the same number of cells. This kind of discretization is called structured grid or simply grid. The major disadvantage of this approach is, that the discretization of highly non-planar elements produces a large number of points in the simulation domain. As a consequence the subsequent simulation and calculation steps are slowed down requiring a lot of computational resources.

The alternative approach is to divide the computational domain into an unstructured assembly of cells. The notable feature of an unstructured mesh is that the number of cells surrounding a typical interior vertex of the mesh is not necessarily constant. This kind of discretization is called unstructured mesh or simply mesh. The major disadvantage of this approach is that the element generation process is one of the most complicated procedures in the field of simulation. However the reduction in simulation time and the requirements on computational resources can be significant.

Based on the complex three-dimensional mesh generation process and the impracticality of using uniform refinement strategies most of the TCAD simulations are based on structured grids. But with the shift to real and complex input structures the grid approach with the involved refinement steps is no longer manageable. Here the unstructured mesh generation techniques come into play. In two dimensions most of the grid or mesh design procedure and adaptation steps are done by hand. With the step from two-dimensional to three-dimensional mesh generation and adaptation a hand crafted design and adaptation is impossible.

First, the user interaction and visualization in three-dimensions is very difficult. Secondly the user can not be aware where the adaptation should be done. On this account three-dimensional mesh generation and adaptation must be coupled with error estimation techniques to ensure an automatic adjustment for a given problem without user interaction.

A difficulty in the field of mesh adaptation is that to this date the understanding of the relationship between the quality of mesh elements, numerical accuracy, and stiffness matrix condition remains incomplete, even for the simplest cases. Experience and mathematical results have shown that isotropic elements usually lead to good results while degenerated elements will negatively affect the computation. Therefore we derive an abstract quality criterion for elements which have to be refined so that automatic remeshing can be easily accomplished by locally removing tetrahedra patches and inserting points derived from the error estimator.

Our novel technique of calculating an abstract quality criterion to control the mesh adaptation or remeshing step separates the mathematical error estimation step from the geometrical meshing step and can therefore be implemented with different error estimation models. Also the software components can be easily upgraded. In the field of unstructured mesh modification the following techniques are possible:

- H-method
  This method uses a geometrical parameter h for refinement (i.e. the height of a tetrahedron).
- P-method
  This method varies the degree p in the approximation (i.e. quadratic ansatz functions within finite elements) while keeping the geometrical size h unchanged.
- HP-method
  This method combines the p-method with the h-method.
- Adaptive remeshing method
  This method extracts a patch of marked elements which are accordingly remeshed.

For our technique we focus mainly on the adaptive remeshing method (some kind of advancing front method [29]) because of the maximum degree of freedom within mesh adaptation.

3.3 Error Estimation

The numerical expression of a discretized problem results in a discrete distribution of quantities and ansatz functions of a certain function class (e.g. piecewise affine functions) to describe the behavior of the quantities. Apart from the quality of the underlying mesh the quality of the simulation essentially depends on the selection of the ansatz functions.

Using piecewise affine or constant ansatz functions like finite volumes or finite elements we always obtain results with a certain error. In terms of function spaces we carry out a projection of the complete space of functions to the subspace of piecewise affine or constant functions. Usually the euclidian norm is used in order to measure the distance between two functions.

$$\| f-g\|_2 = \sqrt{\int_{-\infty}^{+\infty} \big(f(x)-g(x)\big)^2 \mathrm{d}x} \qquad (32)$$

3.3.1 Residual Based Error Estimation

On each triangle the solution function is interpolated piecewise (Figure 16) affinely so as to receive a globally continuous function. This function fulfills the Laplace equation in the interior of the triangle whereas the discontinuity of the interpolated solution function at the boundaries leads to an error which can be estimated locally by the following formula:

$$\eta_k = h_k\Big(\sum_{E\in E_K\cap E_{\mathrm{int}}} \| J_{E,n}(u_h)\|_E^2 + \sum_{E\in E_K} \| J_{E,t}(u_h)\|_E^2\Big) \qquad (33)$$

where E_K denotes the edges of the triangle and E_{int} is the set of the interior edges. The local discontinuity of the gradient of the interpolated function at an edge is $\vec{J_E}$, where $J_{E,n}$ is the normal component and $J_{E,t}$ is the tangential component. The

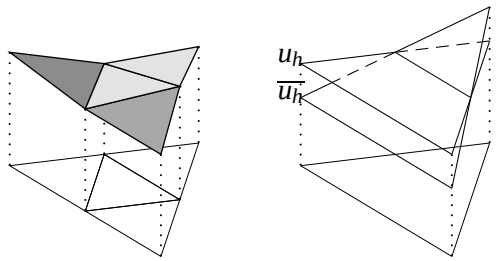


Figure 16: (left) Two-dimensional representation of the error estimator. The normal component of the error changes at the facet. (right) Discrete solution function u_h and the interpolation function $\overline{u_h}$ as function over the mesh triangle

geometry factor h_K denotes a characteristic length of the triangle such as the mean edge length or the circumference radius. An interpretation of the behavior of the error estimator is given in the following. A gradient in the potential causes a flux, which is free of sources in the case of the Laplace equation.

If the flux is discontinuous through a facet of a tetrahedron there has to include source density on the facet. The Laplace equation states, however, that the source density vanishes. Therefore the estimated error is zero if the potential behaves smoothly when crossing a facet. As we use piecewise affine interpolation the function is continous and therefore the jump of the tangential field strength has to vanish. For this reason only the normal components of the field strength are relevant.

3.3.2 ZZ Error Estimation

The ZZ error estimator [30] measures how much the numerical solution u_h differs from a smoothed numerical solution $\overline{u_h}$ (Figure 16). For some types of differential equations such as the Laplace equation the ZZ estimator has been shown to have upper and lower bounds [30].

For the interpolation function of the discrete numeric solution u_h we use polynomial functions of degree one in each tetrahedron. The distance between the interpolated piecewise affine function and the piecewise constant function can be determined by the evaluation of the norm (32) and yields,

$$\eta_k = \sum_i U_i^2 - \sum_{i \neq j} U_i U_j \tag{34}$$

where the U_i are the result values in the vertices of the tetrahedron.

3.3.3 Evaluation of the Error Estimation

A quality statement regarding the calculation can be given counting the simplices within a certain error interval of error values. The range of errors (from zero to the maximum error) is divided into equidistant error classes, i.e. ten classes.

With this separation different adaptation strategies can be used: *minimum number of elements*, *error in element*, or *maximum number of elements*. Here we use the *maximum number of elements* strategy, which is bound to 30% in each refinement step, and the *error in element* strategy, only to sort the elements.

3.4 Results

In the following the results of the error estimation and mesh adaptation techniques are shown. We use two different examples to demonstrate the behavior of our novel technique of coupling the error estimation and mesh adaptation steps through an abstract interface.

First, we use a non-planar capacitor structure and calculate the potential distribution between the contacts. Here we use the residual error estimation technique only to show the shift of the quality of the elements within each error class.

The second example deals with a realistic interconnect line with tapered line elements (lines with angular side walls) and a pyramid element for the via, which connects the two lines. Here we compare the residual error and the ZZ error estimation technique.

For the non-planar capacitor we give a comparison of the initial error and the error value after one remeshing step:

	Initial meshing	Remeshing
Tetrahedra	2,145	8,774
Minimum error	0.02	0.001
Maximum error	25.0	19.4

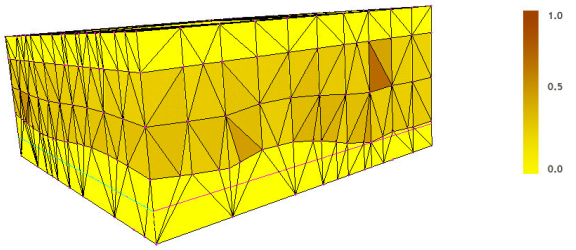


Figure 17: Initial local error values without refinement

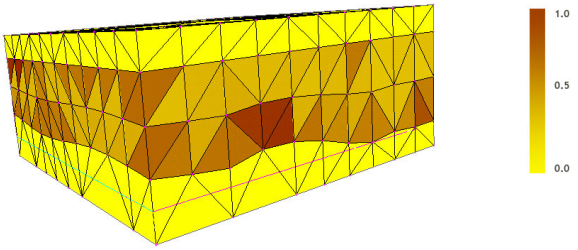
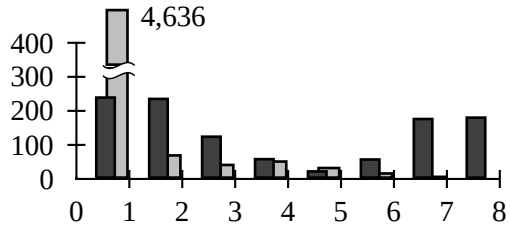


Figure 18: Local error values after one refinement step

The next diagram shows the distribution of error values. The number of tetrahedra is plotted on the y-axis while the x-axis shows the error classes. The light gray boxes show the refined error values whereas the dark grey boxes show the initial error values:



As can be seen, the error values for the elements are shifted to the left side indicating that the local error values drop due to our refinement technique. Figure 17 depicts the error values without any refinement, whereas Figure 18 shows the distribution of error values after one adaptation step (zero stands for a lower error, and one denotes a higher error). As we have seen in the error value diagrams the local error values are shifted to lower values. To show the applicability and usability for a realistic example we solve the Laplace equation within an interconnect structure and show the successfully application of our technique. In Figure 19 we depict the structure, the contacts and the potential distribution. Figure 20 presents a three-dimensional visualization (not a cut through the structure) of the relative error based on the residual error estimation technique within each adaptation step. Compared to the residual error estimator, Figure 21 presents the adaptation steps based on the ZZ error estimation technique. The

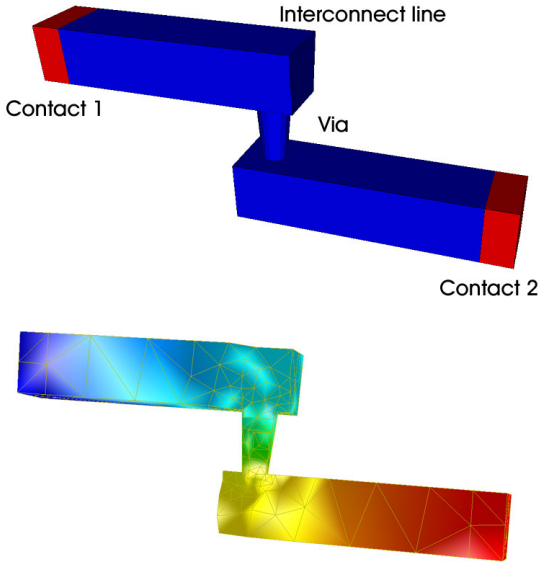


Figure 19: Initial interconnect structure (top) and potential distribution (bottom)

following table gives a comparison of the number of tetrahedra after each adaptation step within the two different error estimation techniques:

	Initial step	Step 1	Step 2
RS: Tetrahedra	1,720	2,052	2,334
ZZ: Tetrahedra	1,720	2,075	2,290

3.5 Conclusion

Using the advantages of mesh refinement in combination with a posteriori error estimation leads to an enormous increase of simulation result quality. The benefits of adaptive mesh refinement allow us

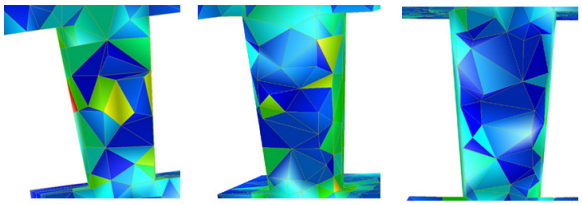


Figure 20: Residual based error estimator, zoomed into the important via: first three adaptation steps

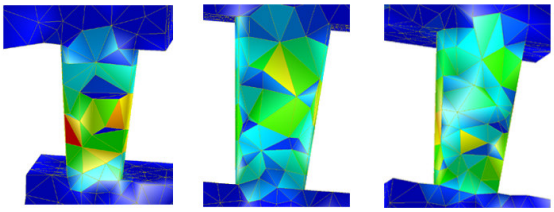


Figure 21: ZZ error estimator, zoomed into the important via: first three adaptation steps

to locally improve the mesh quality without increasing the number of mesh points dramatically. For this reason the resolution of the critical simulation domain is much higher and the relevant processes can be better simulated whereas the regions of lower interest do not require much simulation time. In combination with a posteriori error estimation a measure was found which triggers the refinement and indicates if the quality of the solution is resolved adequately.

4 Level Set Method Based Topography Simulator and its Applications in Interconnect Processes

The application of level set and fast marching methods to the fast simulation of surface topography especially in three dimensions for semiconductor processes is presented. Our general purpose topography simulator is based on these methods and has been implemented using many techniques for increasing of its speed.

4.1 Introduction

Interconnects are becoming increasingly important with shrinking technologies. Capacitance and resistance of interconnect lines determine the timing delays due to metal lines, which contribute more and more to the overall delays. For proper modeling the capacitances, one has to know the metal profile, e.g., bottom and top *CD* (Critical Dimensions) and metal slope, the profile of the deposited layer with and without *CMP* (Chemical Mechanical Planarization), and the profile of the void, if it is formed. These profiles depend heavily on deposition process conditions, metal thickness, and line-to-line spaces, and less strongly on the metal width. The significant influence of void formation in a controlled and reproducible manner as an economically advantage for substituting expensive low-k materials was studied and simulated in two dimensions using our topography simulator [31].

The availability of a fast topography simulator which generates the metal profile and the profile of deposited layers using simulation of etching and deposition processes, is very important. However, three-dimensional topography simulation still faces many challenges which limit its general applicability and usefulness. In addition, three-dimensional topography simulation tends to be very *CPU* and memory expensive to date.

Based on an efficient and precise level set method including narrow banding and extending the speed function in a sophisticated algorithm, we have developed a general topography simulator in two and three dimensions for the simulation of deposition and etching processes. The simulator is called *ELSA* (Enhanced Level Set Application) and works efficiently concerning computational time and memory consumption. It ensures simultaneously high resolution. Furthermore we have developed *TOPO3D* whose kernel is based on *ELSA*, but in addition, it is linked to a program library for handling objects for full three-dimensional semiconductor process simulations. This program library is called WAFER-STATE SERVER.

The outline of this paper is as follows. First, we present briefly the level set method and related techniques for an efficient implementation. Second, we present shortly some two-dimensional simulation results for the backend of a 100*nm* process. Third, some simulation results of three-dimensional structures applicable in interconnect processes are shown. Finally, we shortly describe the WAFER-STATE SERVER.

4.2 The level set method

The level set method provides means for describing boundaries, i.e., curves, surfaces or hypersurfaces in arbitrary dimensions, and their evolution in time which is caused by forces or fluxes normal to the surface [32, 33]. The basic idea is to view the curve or surface in question at a certain time t as the zero level set (with respect to the space variables) of a certain function $u(t,\mathbf{x})$, the so called level set function. Thus the initial surface is the set $\{\mathbf{x} \mid u(0,\mathbf{x}) = 0\}$.

Each point on the surface is moved with a certain speed normal to the surface and this determines the time evolution of the surface. The speed normal to the surface will be denoted by $F(t,\mathbf{x})$.

The surface at a later time t_1 shall also be considered as the zero level set of the function $u(t,\mathbf{x})$, namely $\{\mathbf{x} \mid u(t_1,\mathbf{x}) = 0\}$. This leads to the level set equation

$$u_t + F(t,\mathbf{x})\|\nabla_{\mathbf{x}} u\| = 0, \qquad u(0,\mathbf{x}) \quad \text{given},$$

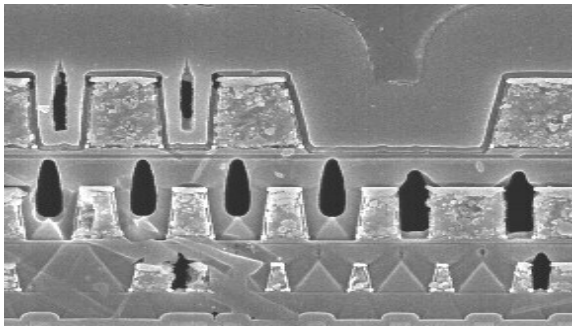


Figure 22: SEM image of a whole backend stack comprised of three Al metals and a Ti-nitride interconnect.

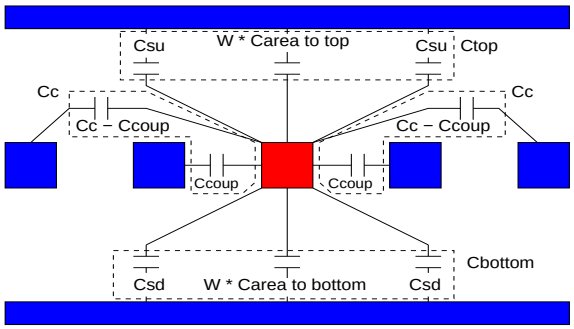


Figure 23: Two-dimensional schematics of a signal line (middle) surrounded by grounded lines and planes.

in the unknown variable u, where $u(0,\mathbf{x})$ determines the initial surface. Having solved this equation the zero level set of the solution is the sought curve or surface at all later times.

Now in order to apply the level set method a suitable initial function $u(0,\mathbf{x})$ has to be determined first. A beneficial choice is the signed distance function of a point from the given surface. After calculation of the initial level set function, the speed function values on the whole grid are used to update the level set function in a finite difference or finite element scheme. Usually the values of the speed function are not determined on the whole domain by the physical models [34, 35] and, therefore, have to be extrapolated suitably from the values provided on the boundary, i.e., the zero level set. This can be carried out iteratively by starting from the points nearest to the surface. The idea leading to fast level set algorithms stems from observing that only the values of the level set function near its zero level set are essential, and thus

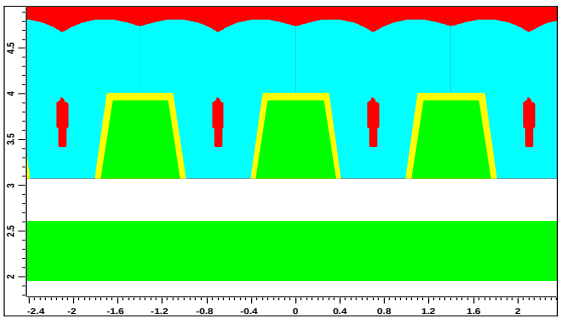


Figure 24: Simulation of void formation by M3 lines at $0.90\mu m$ space above the M2 plane.

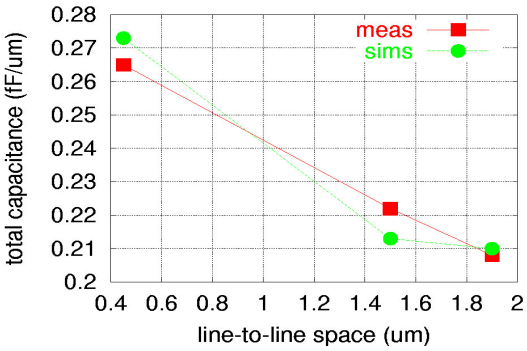


Figure 25: Comparison between simulation and measurement of M3 middle line capacitance as a function of line-to-line spaces.

only the values at the grid points in a narrow band around the zero level set have to be calculated. Both improvements, extending the speed function and narrow banding, require the construction of the distance function from the zero level set in the order of increasing distance. But calculating the exact distance function from a surface consisting of a large number of small triangles is computationally expensive and can be only justified for the initialization. An approximation to the distance function can be computed by a special fast marching method [36, 37].

4.3 Two-Dimensional Interconnect Capacitance Simulation

In the process considered the films deposited as *ILD* (Interlayer Dielectric) are silicon nitride and silicon dioxide films. For topography simulation

of a deposition process, it is generally possible to consider complicated reaction paths. However, it is advantageous to reduce the possible reaction to an essential minimum for reducing the complexity of the simulator. Figure 22 shows a whole backend stack comprised of three different metal lines M1, M2, and M3, bottom-up, respectively.

In order to model capacitances, we assume that a signal line is at high voltage and surrounded by two lines on the left, two lines on the right, a plane underneath, and a plane above. The surrounding lines and planes are assumed to be at ground voltage. For example an M2 signal line could be surrouded by M2 grounded lines above the M1 plane and underneath the M3 plane. This skeleton is shown in Figure 23. Most resistance and capacitance extraction *RCX* tools have very simplistic void models. Even if the metal slope is modeled, it is mostly assumed constant and independent of space. This is insufficient for today's technologies where interconnects have a large number of special features which are nowhere close to ideal [38, 39].

Figure 24 shows the simulation result of void formation at $90\mu m$ line-to-line spaces where the slopes of metal lines have been assumed to dependent on line-to-line spaces. A very good agreement between the simulations and measurements of M3 line capacitances with an error of less than 5% has been achieved as shown in Figure 25.

4.4 Three-Dimensional Simulation Results

In this section we present three-dimensional simulation results obtained by *ELSA* for detection of void formation in interconnect lines after deposition of *ILD* materials. Figure 26 shows the three-dimensional structure, with metal width W, line-to-line spaces S, and metal thickness T in x, y, and z direction, labeled with X, Y, and Z, respectively. The deposited layers were silicon dioxide and silicon nitride with thickness of $D1 = 0.1\mu m$ and $D2 = 0.9\mu m$, respectively.

The goal of simulation was detecting the void for a set of different S holding the metal thickness

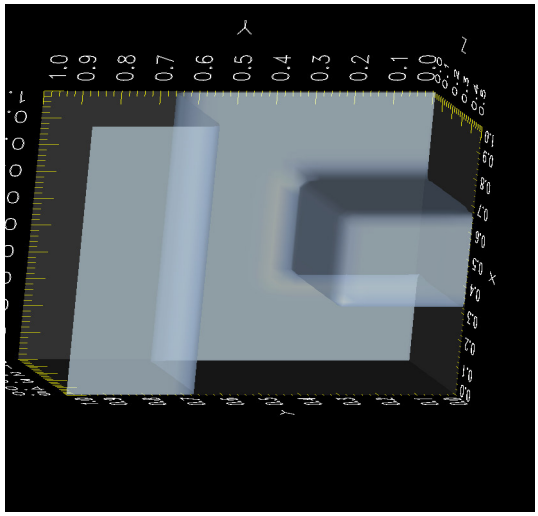


Figure 26: Three-dimensional initial boundary for the deposition of silicon dioxide and silicon nitride in an interconnect structure.

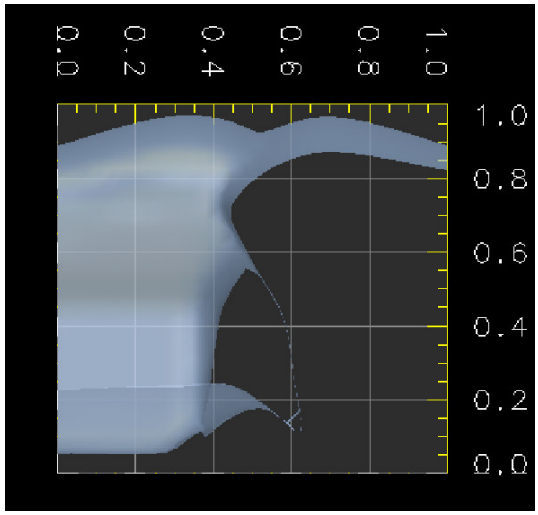


Figure 27: A y-z point of view of simulation of void formation at $S = 0.72\mu m$ using *ELSA*.

one time at $T1 = 1.045\mu m$, and for second time at $T2 = 0.845\mu m$. We introduce a parameter C(T,S) which is calculated as follows:

$$C(S,T) = T + D1 + D2 - H_{void}(S)$$

where H_{void} stands for the z coordinate of the top of a void. In order to know how C(T,S) depends on line-to-line spaces and metal thickness, we have performed four different simulations at $S = 0.18, 0.36, 0.72,$ and $1\mu m$, for $T1$ and for $T2$. Figure 27 shows a y-z point of view of the

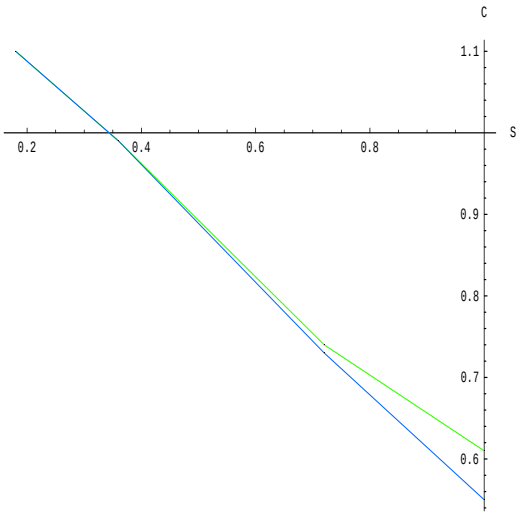


Figure 28: Dependence of C on S for $T1 = 1.045\mu m$, and $T2 = 0.845\mu m$. The lower and upper curve stands for $T2$ and $T1$, respectively.

three-dimensional simulation of void formation of the initial structure shown in Figure 26 for $S = 0.72\mu m$ and $T1$.

Figure 28 shows the simulation result for calculating of C for different line-to-line-spaces at $T1$, and $T2$. As expected, the z coordinate of top of void has been greater as we have increased S. Whereas for small line-to-line spaces the metal thickness does not play an important role, the effect of metal thickness will be more important with increasing S. Furthermore, the simulation results have shown that the metal width can not considerably affect the void formation and its dimensions.

4.5 Wafer State Server

The WAFER-STATE SERVER is a program library and file format for handling three-dimensional objects in semiconductor process simulation developed at the Institute for Microelectronics [40]. It is a solution to the integrated simulation of three-dimensional manufacturing processes.

A generic data model suitable for process and device simulations allows for an efficient data exchange between simulators even when they are based on different native file formats. It is able to handle different meshes and distributed quantities

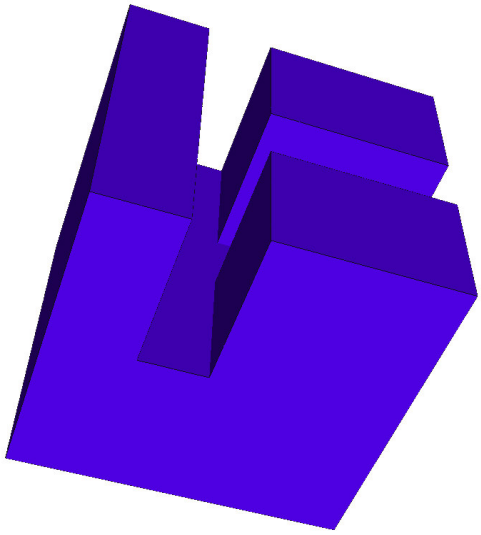


Figure 29: A T-shape initial boundary for an isotropic deposition process.

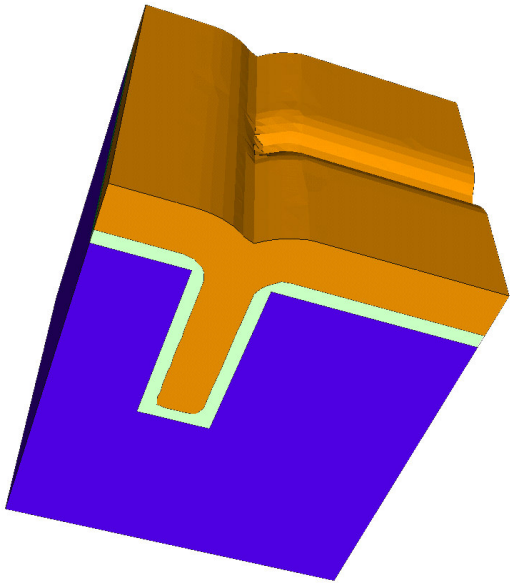


Figure 30: Simulation result of isotropic deposition of two different materials using *TOPO3D*.

stored thereon. The program library also defines algorithms to perform geometrical operations for full three-dimensional process simulation as they are used in topography simulations.

Figure 30 shows the simulation result of an isotropic deposition into a T-shape initial boundary shown in Figure 29 using *TOPO3D*. The deposition simulation have been done for two different materials.

4.6 Conclusion

State of the art algorithms for surface evolution processes like deposition and etching processes in three dimensions have been implemented. A general purpose topography simulator was developed based on the level set method combining narrow banding and fast marching methods for extending the speed function. The application of the simulator has been presented for two and three-dimensional interconnect processes. A very good agreement between the two-dimensional simulation results and measurements of capacitance after deposition of *ILD* materials has been achieved. A set of three-dimensional simulations for different line-to-line spaces and metal thicknesses has been performed. Line-to-line spaces and metal thickness play an important role by void formation and its dimensions. However, the effect of metal thickness on void profile will only be important from a determined line-to-line spaces.

5 Simulation of Dynamic NBTI Degradation for a 90nm CMOS Technology

The NBTI effect has become the limiting factor for the reliability of p-MOSFETs in the sub-100nm regime. In this work the dynamic NBTI degradation was systematically investigated for a 90nm p-MOSFET by experiment and simulation. For thin gate oxides stressed at low to medium gate voltages the bulk traps can be neglected and NBTI occurs mainly due to the generation of interface traps. Under this condition the reaction-diffusion model can be applied for the prediction of NBTI degradation. The model parameters were calibrated for NBTI simulations at arbitrary gate voltage, frequency, and duty cycle within a calibrated range. The long-time NBTI degradation was simulated up to 10 years in order to estimate the transistor lifetime under typical chip operation conditions.

5.1 Introduction

Negative bias temperature instability (NBTI) leads to rapid negative shifts of the p-MOSFET threshold voltage V_T due to the buildup of positive charged interface traps N_{it}. The generation of interface traps occurs when the transistor is stressed by negative gate voltages at elevated temperatures. NBTI degradation increases the absolute V_T value and reduces the drain current and transconductance of the transistor. It was found that the DC degradation follows a power-law time dependence which can be approximately described by $N_{it}(t) \propto t^{\frac{1}{4}}$. Under AC operation the interface traps generated during the on-state of the transistor are partially annealed in the off-state. Compared to static NBTI behavior the dynamic NBTI effect thus significantly improves the transistor lifetime.

While the microscopic details of the NBTI mechanism are not fully understood, it is speculated that interface traps are generated by dissociation of Si–H bonds at the silicon-oxide interface during NBTI stress [41]. A released hydrogen atom diffuses away from the silicon-oxide interface and leaves an interface trap behind which is charged positive. Although interface traps are mainly responsible for NBTI in the majority of the cases, oxide traps dominate the overall degradation for high gate voltages particularly in thick oxide devices [42].

In order to allow the treatment of bulk traps as well, an equivalent positive sheet charge located at the silicon-oxide interface can be defined, which is caused by interface and bulk traps during inversion of the p-MOSFET. The equivalent sheet charge density ($N_{it} \cdot q$) produces a threshold voltage shift ΔV_T which depends on the oxide capacitance per unit area, C_{ox}, according to

$$\Delta V_T = -\frac{N_{it} \cdot q}{C_{ox}} \tag{35}$$

The V_T degradation in the p-MOSFET reduces the gate overdrive ($V_G - V_T$) wich leads to a reduced drive current. This is consistent with the observation that the rise time of the CMOS inverter output signal, controlled by the p-MOSFET, degrades over time, whereas the fall time, controlled by the n-MOSFET, stays unchanged. The magnitude of the NBTI induced parameter shift depends also significantly on the frequencies and duty cycles which occur during operation of the transistor [43].

Nitrogen plays a key role in NBTI sensitivity, especially if it is located near the silicon-oxide interface. It can be speculated that the application of thinner SiON gate dielectrics leads to a faster hydrogen diffusion due to the lower quality of the nitrided oxide compared to pure silicon dioxide. The faster loss of hydrogen at the interface increases the NBTI degradation. For heavily nitrided, thin gate oxides at low electric fields NBTI may dominate over hot carrier injection (HCI) which occurs primarily in the n-MOSFET [44].

In this paper NBTI degradation was systematically investigated for the p-MOSFET of a 90nm technology with 20Å equivalent oxide thickness (EOT). Specific experiments were performed in order to analyze the gate voltage, duty cycle, and frequency dependence of the NBTI degradation behavior.

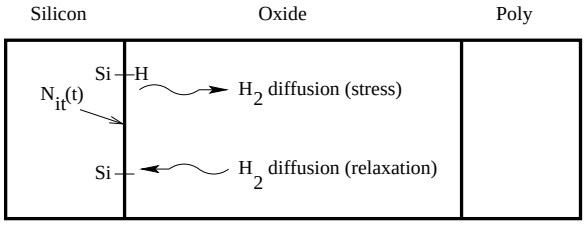


Figure 31: Description of the R-D model [4].

Frequency measurements were performed in the range from DC to 1 MHz and "ON" duty cycles in the range of 30% to 70% were used. Stress voltages were applied from -1.5 V up to -2.7 V to the gate of the transistor at a temperature of 125° C. The collected measurement data were used to fit the model parameters for NBTI simulations.

5.2 Reaction-Diffusion Model

A reaction-diffusion (R-D) model is used for the simulation of NBTI degradation, because this model accurately reproduces experimental NBTI data [45]. The R-D model states that the buildup of interface traps N_{it} arises from dissociation of hydrogen (constant dissociation rate k_f) governed by an electrochemical reaction between inversion layer holes and Si–H bonds.

The released hydrogen diffuses away from the silicon-oxide interface or reacts with a dangling silicon bond. The N_{it} passivation occurs with anneal rate k_r [41]. When the transistor is switched off, annealing occurs at unchanged anneal rate k_r and dissociation rate $k_f = 0$.

The R-D model is schematically explained in Figure 31. This sketch demonstrates that the released hydrogen diffuses into the oxide during the stress phase and returns to the interface during the relaxation phase. In the first few seconds the trap generation is controlled by the reaction process (fast N_{it} build-up), while the long-time generation is governed by the diffusion process (slow N_{it} generation).

The stress phase (N_{it} generation) and relaxation phase (N_{it} annealing) can be observed in Figure 32. The R-D model is described by the following two coupled differential equations

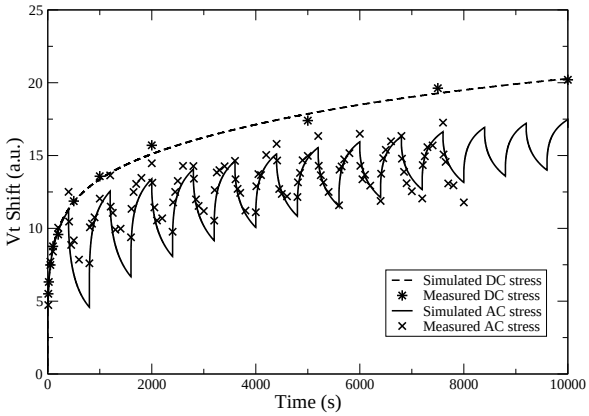


Figure 32: Simulated and measured NBTI.

$$\frac{\partial N_{it}(t)}{\partial t} = k_f\,[N_0 - N_{it}(t)] - k_r\,N_{it}(t)\,C_H(x,t)\,|_{x=0} \quad (36)$$

$$\frac{\partial N_{it}(t)}{\partial t} = -\,D\,\frac{\partial C_H(x,t)}{\partial x}\Big|_{x=0} + \frac{\delta}{2}\frac{\partial C_H(x,t)}{\partial t} \quad (37)$$

A prerequisite for the applicability of the R-D model is that the generated bulk traps during NBTI aging can be neglected compared to the generated interface defect density. Recently, charge pumping and stress-induced leakage current (SILC) measurements revealed that this prerequisite is fulfilled for thinner oxides stressed at low to medium gate voltages [42].

Also, the generation of bulk traps caused by injection of hot holes into the oxide shows a much stronger voltage dependence than the interface trap generation. However, an excellent ΔV_T versus N_{it} correlation over a wide range of gate voltages supports the absence of bulk traps and that ΔV_T is purely due to interface traps N_{it} [42].

5.3 Simulation

A one-dimensional finite differences method was applied for the discretization of the differential equations with the boundary condition of an absorbing wall at the oxide-poly interface. A neutral diffusion species (atomic or molecular hydrogen) is assumed and the hydrogen distribution profile in the oxide $C_H(x,t)$ is calculated for every time step [46].

The result is the defect density N_{it} and the corresponding shift $\Delta V_T \propto N_{it}$. Figure 32 shows the

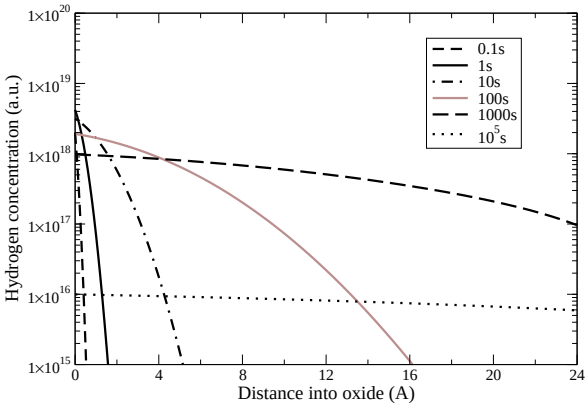


Figure 33: Hydrogen diffusion profiles.

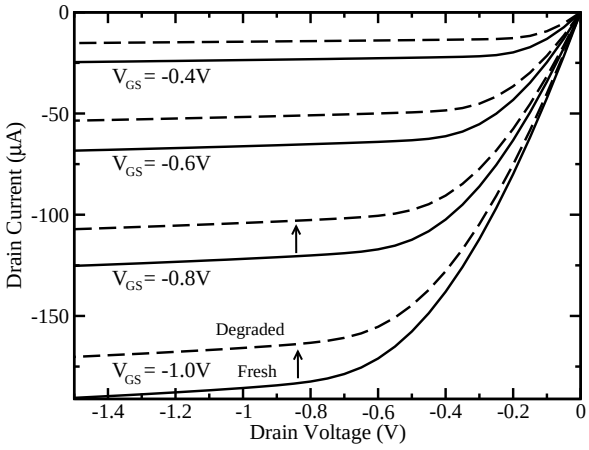


Figure 34: Degraded output characteristics.

simulation result for DC stress compared to some cycles of AC stress at a very low frequency. It demonstrates a good agreement of the predicted V_T degradation with measured NBTI data.

Figure 33 shows snap-shots of the corresponding hydrogen distribution in the oxide during DC stress. Note that the traps at the interface are built up quickly during the first seconds (reaction-limited regime) corresponding to a fast generation of a high level of hydrogen concentration in the interface zone. With increasing time the hydrogen diffusion front moves towards the oxide-poly interface (diffusion-limited regime). After long times the hydrogen concentration at the silicon-oxide interface becomes low again corresponding to a high level of generated defects.

The calculated N_{it} value can further be used as input value for the device simulator Minimos-NT in order to simulate the degraded output characteristics of the p-MOSFET. Figure 34 shows the Minimos-NT result for the degraded drain curent of a p-MOSFET with 30Å EOT at the end of the lifetime, defined by $\Delta V_T = 60mV$ [10].

5.3.1 Gate Voltage Dependence

In order to allow an NBTI simulation under more realistic operation conditions, the gate voltage dependence was included in the simulations. The dissociation rate k_f is determined by the available surface hole concentration which depends on the applied gate voltage. The dissociation rate is directly proportional to the inversion hole density P ($k_f \propto P$) [42]. The inversion holes tunnel to the Si–H bonds located in the SiO_x interface zone governed by the electrical field.

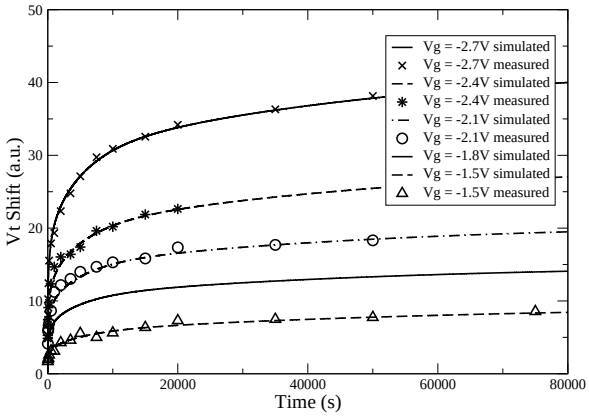


Figure 35: NBTI gate voltage dependence.

The amount of Si–H bonds N_0 at the interface which can be reached by this stochastic tunneling process increases with the oxide field. The holes get captured and take away one electron from the Si-H bonds. The weakened Si-H bonds are then broken by thermal excitation [42].

Figure 35 shows an excellent agreement between the simulation results and NBTI experimental data in the whole measurement range from -1.5V up to -2.7V. It can be observed that the V_T degradation depends on the gate voltage in a non-linear manner, especially in the higher stress voltage regime.

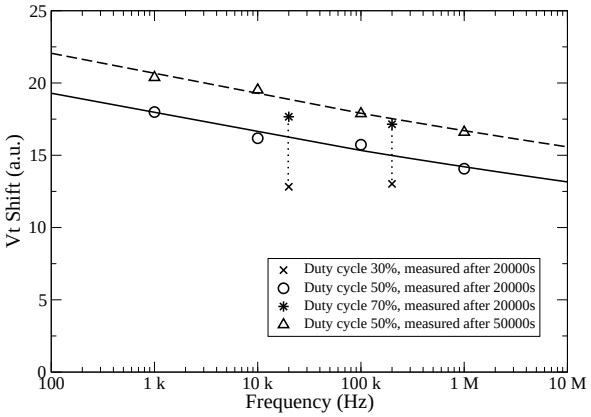


Figure 36: NBTI frequency dependence.

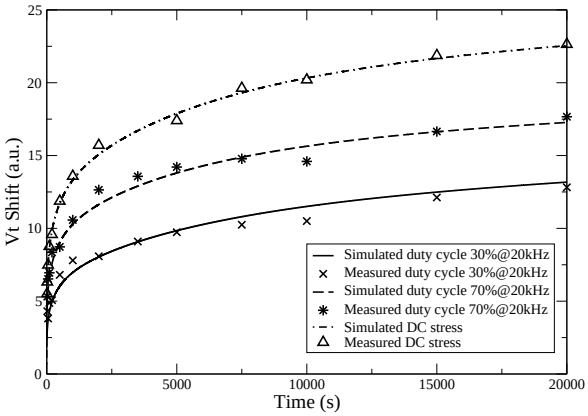


Figure 37: NBTI duty cycle dependence.

5.3.2 Frequency Dependence

Contrary statements can be found in the literature, whether NBTI depends on the frequency or not. No frequency dependence was found for NBTI measurements up to 200kHz after a stress time of only 1000s [47]. However, as depicted in Figure 36 the NBTI degradation is significantly reduced for higher frequencies and/or smaller "ON" duty cycles. The measured frequency dependence can be described by using a reference frequency $f_0 \in [1\text{kHz},1\text{MHz}]$ according to

$$V_T(f) = V_T(f_0) \left(\frac{f}{f_0}\right)^{-0.03323} \tag{38}$$

The reaction-diffusion model predicts no frequency dependence of the NBTI degradation. In this first order model the dynamic NBTI degradation depends only on the duty cycle of the gate signal.

On the other side, NBTI simulations in the range of years for high frequency operation are not feasible due to the required small time resolution which should be in the range of about $\frac{1}{10}$ of the period duration.

For fast simulation at high frequencies we suggest to perform the simulation at low frequency f_0 with equal duty cycle and to correct the result afterwards with the NBTI shift as function of the frequency according to (4). As depicted in Figure 38, this simple approach can accurately predict measurements for 20 kHz at different duty cycles.

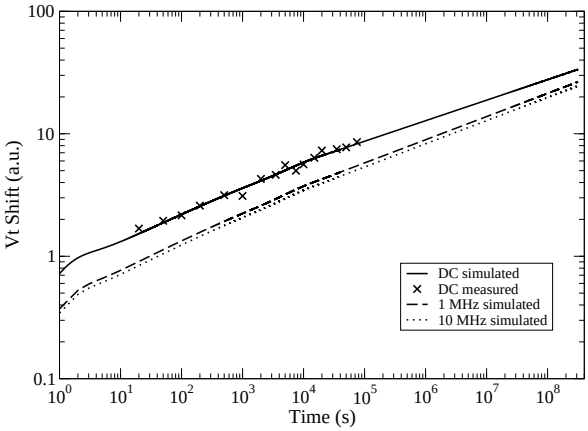


Figure 38: Long-time NBTI simulation for 10 years.

5.3.3 Lifetime Estimation

Currently there is no agreed standard procedure available for the characterization and the measurement of NBTI reliability. Companies use different failure criterions for the definition of the device lifetime. Often the stress conditions for the NBTI measurements are chosen in a way which allows to extrapolate a theoretical transistor lifetime of 10 years.

A reasonable failure criterion for the lifetime of the 90nm technology is, for instance, $\Delta V_T/V_T = 10\%$. For other transistor types, such as power devices, this criterion may be too strict since a higher V_T shift may be tolerable. Figure 34 compares long-time NBTI degradation simulations for the lifetime estimation under DC

and typical chip operation frequencies. It can be observed that the diffusion-limited regime is reached after about two seconds which is characterized by a reduced slope of the V_T shift over time.

5.4 Conclusion

The NBTI mechanism was systematically investigated for a 90nm CMOS technology. Experiments at different gate voltages, frequencies, and duty cycles were performed in order to analyze the NBTI behavior of the p-MOSFET. The simulation method is based on the numerical solution of the reaction-diffusion model. The R-D model was extended to include the gate voltage and frequency dependence. The successful calibration of the model parameters is demonstrated by comparing the simulation results with measured NBTI data. All experimental data could be well reproduced. The presented simulation approach allows to predict the p-MOSFET lifetime depending on the applied stress operation conditions.

References

[1] A. Gehring and S. Selberherr. Modeling of Tunneling Current and Gate Dielectric Reliability for Nonvolatile Memory Devices. *IEEE Trans.Device and Materials Reliability*, 4(3):306–319, 2004.

[2] W. J. Chang, M. P. Houng, and Y. H. Wang. Simulation of Stress-Induced Leakage Current in Silicon Dioxides: A Modified Trap-Assisted Tunneling Model considering Gaussian-Distributed Traps and Electron Energy Loss. *J.Appl.Phys.*, 89(11):6285–6293, 2001.

[3] F. Jiménez-Molinos, A. Palma, F. Gámiz, J. Banqueri, and J. A. Lopez-Villanueva. Physical Model for Trap-Assisted Inelastic Tunneling in Metal-Oxide-Semiconductor Structures. *J.Appl.Phys.*, 90(7):3396–3404, 2001.

[4] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli. Modeling of Anomalous SILC in Flash Memories Based on Tunneling at Multiple Defects. *Solid-State Electron.*, 46(11):1749–1756, 2002.

[5] L. Larcher. Statistical Simulation of Leakage Currents in MOS and Flash Memory Devices with a New Multiphonon Trap-Assisted Tunneling Model. *IEEE Trans.Electron Devices*, 50(5):1246–1253, 2003.

[6] E. Rosenbaum and L. F. Register. Mechanism of Stress-Induced Leakage Current in MOS Capacitors. *IEEE Trans.Electron Devices*, 44(2):317–323, 1997.

[7] F. Schuler, R. Degraeve, P. Hendrickx, and D. Wellekens. Physical Description of Anomalous Charge Loss in Floating Gate Based NVM's and Identification of its Dominant Parameter. In *Intl. Reliability Physics Symposium*, pp 26–33, 2002.

[8] A. Palma, A. Godoy, J. A. Jimenez-Tejada, J. E. Carceller, and J. A. Lopez-Villanueva. Quantum Two-Dimensional Calculation of Time Constants of Random Telegraph Signals in Metal-Oxide-Semiconductor Structures. *Physical Review B*, 56(15):9565–9574, 1997.

[9] M. Herrmann and A. Schenk. Field and High-Temperature Dependence of the Long Term Charge Loss in Erasable Programmable Read Only Memories: Measurements and Modeling. *J.Appl.Phys.*, 77(9):4522–4540, 1995.

[10] Institut für Mikroelektronik, Technische Universität Wien, Austria. *MINIMOS-NT 2.1 User's Guide*, 2004. `www.iue.tuwien.ac.at/software`.

[11] D. R. Hamann. Diffusion of Atomic Oxygen in SiO_2. *Phys. Rev. Lett*, 81(16):3447–3450, 1998.

[12] Ch. Hollauer, H. Ceric, and S. Selberherr. Simulation of Thermal Oxidation: A Three-Dimensional Finite Element Approach. In *Proc. of the 33rd European Solid State Device Research Conference (ESSDERC 2003)*, pp 383–386, Estoril, Portugal, 2003.

[13] E. Rank and U. Weinert. A Simulation System for Diffuse Oxidation of Silicon: A Two-Dimensional Finite Element Approach. *IEEE Trans. on CAD*, 9(5):543–550, 1990.

[14] B. E. Deal and A. S. Grove. General Relationship for the Thermal Oxidation of Silicon. *J.Appl.Phys.*, 36(12):3770–3778, 1965.

[15] A. Poncet. Finite-Element Simulation of Local Oxidation of Silicon. *IEEE Trans. on CAD*, 4(1):41–53, 1985.

[16] S. Cea. *Multidimensional Viscoelastic Modelling of Silicon Oxidation and Titanium Silicidation*. PhD thesis, University of Florida, USA, 1996.

[17] V. Senez, S. Bozek, and B. Baccus. 3-Dimensional Simulation of Thermal Diffusion and Oxidation Processes. In *Proc. International Electron Devices Meeting (IEDM 1996)*, pp 705–708, San Francisco, USA, 1996.

[18] H. Z. Massoud and J. D. Plummer. Analytic Relationship for the Oxidation of Silicon in Dry Oxygen in the Thin-Film Regime. *J.Appl.Phys.*, 62(8):3416–3423, 1987.

[19] H. Z. Massoud, J. D. Plummer, and E. A. Irene. Thermal Oxidation of Silicon in Dry Oxygen. *J. Electrochem. Soc.*, 132(7):1746–1753, 1985.

[20] H. Z. Massoud, J. D. Plummer, and E. A. Irene. Thermal Oxidation of Silicon in Dry Oxygen Growth-Rate Enhancement in the Thin Regime. *J. Electrochem. Soc.*, 132(11):2685–2693, 1985.

[21] A. J. Moulson and J. P. Roberts. Water in Silica Glass . *J. Trans. Farady Soc.*, 57:1208–1216, 1961.

[22] F. J. Norton. Permeation of Gaseous Oxygen through Vitreous Silica. *Nature*, 191:701, 1961.

[23] V. Senez, D. Collard, P. Ferreira, and B. Baccus. Two-Dimensional Simulation of Local Oxidation of Silicon: Calibrated Viscoelastic Flow Analysis. *IEEE Trans.Electron Devices*, 43(5):720–731, 1996.

[24] H. Matsumoto and M. Fukuma. A Two-Dimensional Si Oxidation Model including Viscoelasticity. In *Proc. International Electron Devices Meeting (IEDM 1983)*, pp 39–43, Washington, USA, 1983.

[25] V. Senez et al. Analysis and Application of a Viscoelastic Model for Silicon Oxidation. *J.Appl.Phys.*, 76(6):3285–3296, 1994.

[26] J. P. Peng, D. Chidambarrao, and G. R. Srinivasan. Novel: A Nonlinear Viscoelastic Model for Thermal Oxidation of Silicon. *COMPEL - The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, 10(4):341–353, 1991.

[27] O. C. Zienkiewicz. *The Finite Element Method: Basic Formulation and Linear Problems*, volume 1. McGraw - Hill, Maidenhead, England, 4 edition, 1987.

[28] S. Prudhomme et al. Practical Methods for a Posteriori Error Estimationin Engineering Applications. *International Journal for Numerical Methods in Engineering*, 56(8):1193–1224, 2003.

[29] P. Fleischmann. Enhanced Advancing Front Delaunay Meshing in TCAD. In *Proc. International Conference on the Simulation of Semiconductor Processes and Devices (SISPAD 2002)*, pp 99–102, Kobe, Japan, 2002.

[30] O. C. Zienkiewicz and J. Z. Zhu. A Simple Error Estimator and Adaptive Procedure for Practical Engineering Analysis. *International Journal for Numerical Methods in Engineering*, 24:337–357, 1987.

[31] C. Heitzinger et al. Feature-Scale process Simulation and Accurate Capacitance Extraction for the Backend of a 100-nm Aluminum/TEOS Process. *IEEE Trans.Electron Devices*, 51(7):1129–1134, 2004.

[32] J. Sethian. *Level Set Methods and Fast Marching Methods*. Cambridge University Press, Cambridge, England, 1999.

[33] C. Heitzinger. *Simulation and Inverse Modeling of Semiconductor Manufacturing Processes*. PhD thesis, Technische Universität, Wien, 2002.

[34] T. S. Cale. Flux Distributions in Low Pressure Deposition and Etch Models. *J. Vac. Sci. Technol. B*, 9:2551–2553, 1991.

[35] T. S. Cale and G. B. Raupp. A Unified Line-of-Sight Model of Deposition in Rectangular Trenches. *J. Vac. Sci. Technol. B*, 8:1242–1248, 1990.

[36] A. Sheikholeslami et al. Three-Dimensional Topography Simulation for Deposition and Etching Processes Using a Level Set Method. In *Proc. of International Conference on Microelectronics (MIEL)*, pp 241–244, Nis, Serbia, 2004.

[37] A. Sheikholeslami et al. Three-Dimensional Topography Simulation Based on a Level Set Method. In *Proc. of Internatonal Spring Seminar on Electronics Technolog (ISSE)*, pp 263–265, Sofia, Bulgaria, 2004.

[38] F. Badrieh et al. From Feature Scale Simulation to Backend Simulation for a 100nm CMOS Process. In *Proc. of the 33rd European Solid State Device Research Conference (ESSDERC 2003)*, pp 441–444, Estoril, Portugal, 2003.

[39] A. Sheikholeslami et al. Simulation of Void Formation in Interconnect Lines. In *Proc. of SPIE's first International Symposium on Microtechnologies for the New Millennium: VLSI Circuit and Systems*, pp 445–452, Gran Canaria, Spain, 2003.

[40] T. Binder. *Rigorous Integration of Semiconductor Process and Device Simulators*. PhD thesis, Technische Universität, Wien, 2002.

[41] M. A. Alam. A Critical Examination of the Mechanics of Dynamic NBTI for PMOSFETs. In *Proc. International Electron Devices Meeting (IEDM 2003)*, pp 345–348, Washington, USA, 2003.

[42] S. Mahapatra, P. B. Kumar, and M. A. Alam. Investigation and Modeling of Interface and Bulk Trap Generation During Neagtive Bias Temperature Instability of p-MOSFETs. *IEEE Trans.Electron Devices*, 51(9):1371–1379, 2004.

[43] H. Puchner and L. Hinh. NBTI Reliability Analysis for a 90nm CMOS Technology. In *Proc. of the 34th European Solid State Device Research Conference (ESSDERC 2004)*, pp 257–260, Leuven, Belgium, 2004.

[44] D. K. Schroder and J. A. Babcock. Negative Bias Temperature Instability: Road to Cross in Deep Submicron Silicon Semiconductor Manufactoring. *J.Appl.Phys.*, 94(1):1–18, 2003.

[45] W. Abadeer and W.Ellis. Behavior of NBTI Under AC Dynamic Circuit Conditions. In *Proc. Int. Reliability Phys. Symp.*, pp 17–22, 2003.

[46] S. Ogawa and N. Shiono. Generalized Diffusion-Reaction Model for the Low-Field Charge-Buildup Instability at the Si–SiO_2 Interface. *Physical Review B*, 51(7):4218–4230, 1995.

[47] G. Chen et al. Dynamic NBTI of PMOS Transistors and Its Impact on Device Lifetime. In *Proc. Int. Reliability Phys. Symp.*, pp 196–202, 2003.