TCAD Modeling of Negative Bias Temperature Instability

T. Grasser°, R. Entner°, O. Triebl°, H. Enichlmair*, and R. Minixhofer*
°Christian Doppler Laboratory for TCAD in Microelectronics
at the Institute for Microelectronics, TU Wien, Gußhausstraße 27–29/E360, 1040 Wien, Austria
Phone: +43-1-58801/36023, Fax: +43-1-58801/36099, E-mail: grasser@iue.tuwien.ac.at
*austriamicrosystems, Tobelbaderstrasse 30, 8141 Unterpremstätten, Austria

***Abstract*— At elevated temperatures, pMOS transistors show a considerable drift in fundamental device parameters such as the threshold voltage when a large negative bias is applied. This phenomenon, known as negative bias temperature instability, is regarded as one of the most important reliability concerns in highly scaled pMOS transistors. Modeling efforts date back to the reaction-diffusion (RD) model proposed by Jeppson and Svensson forty years ago which has been continuously refined since then. So far, the change in the interface state density predicted by the RD model is directly used to approximate the threshold voltage shift. Here we present a coupling of the RD model to the semiconductor equations which is required to go beyond that approximation and to study degradation during realistic device operating conditions. It is also shown that such a coupled treatment is required to accurately model the behavior during the measurement phase. In addition, the RD model is extended to improve the prediction both in the stress and the relaxation phase by accounting for trap-controlled transport of the released hydrogen species.**

I. Introduction

Since its discovery forty years ago [1] negative bias temperature instability (NBTI) has received a lot of scientific attention [2–6]. Due to the increased electric fields inside the gate-oxide, the presence of nitrogen, and the increased operating temperatures, NBTI has become one of the most important reliability concerns in modern CMOS technology [2, 6]. Negative bias temperature stress is normally introduced via a large negative voltage at the gate with drain and source remaining grounded and results in a shift in device parameters, for instance in the threshold voltage, the subthreshold slope, and the mobility. Most importantly, the shift of the threshold voltage is often described by a simple power-law $\Delta V_{\rm th}(t) = At^n$, with A being a coefficient which depends on temperature and the electric field. While in earlier investigations the exponent n was given to be in the range $0.2 - 0.3$, newer investigations show that n can be as small as 0.12. In particular it was found that the experimentally determined exponent is very sensitive to the measurement setup, most notably to the unintentional delay introduced thereby.

NBTI degradation is most commonly modeled by some form of reaction-diffusion (RD) model which is often solved analytically [3, 7, 8], and sometimes numerically in one or two dimensions [3, 7, 9, 10]. Since the forward rate of the RD model depends on the electric field and the created interface states influence the electric field distribution, a rigorous study requires the coupling of the RD model to the semiconductor equations, for instance the drift-diffusion model. Although the computational effort increases considerably, the advantages of such a coupled approach are manifold. First, the feedback of the various charges on the RD model is considered and, most importantly, $V_{\rm th}$ shifts can be extracted by mimicking the dynamics of the measurement process (stress interruption, unavoidable relaxation, $V_{\rm G}$ sweep, and $I_{\rm D}$ monitoring) instead of relying on simplified relations. Furthermore, such a coupled approach allows to study device degradation under realistic dynamic and even inhomogeneous boundary conditions as they occur under device operation. However, such a coupling requires some care and some important considerations are discussed in the following.

II. The Reaction-Diffusion Model

The reaction-diffusion model dates back to the work of Jeppson and Svensson [11] and has been continuously refined [3, 4, 7, 9]. It relies on the depassivation of dangling bonds at the Si/SiO_2 interface during stress. These dangling bonds, which are commonly known as $P_{\rm b}$ centers [12, 13], are present in a considerable number at every Si/SiO_2 interface and have to be passivated during device fabrication by some sort of hydrogen anneal [2]. At elevated electric fields and temperatures the hydrogen-silicon bonds can be broken, thereby releasing atomic hydrogen. The released H either diffuses through the oxide, resulting in a slope of $n = 1/4$ or quickly dimerizes into H_2 [3, 4, 14], which gives a slope of $n = 1/6$.

The kinetic equation describing the interface reaction is [7]

$$\frac{\partial N_{\rm it}}{\partial t} = k_{\rm f}(N_0 - N_{\rm it}) - k_{\rm r} N_{\rm it} H_{\rm it}^{1/a} \ , \qquad (1)$$

where $N_{\rm it} = [P_{\rm b}]$ is the interface state density, $N_0 = [P_{\rm b}{\rm H}]_0$ the initial density of passivated interface defects, $H_{\rm it}$ the interfacial hydrogen concentration, $k_{\rm f}$ and $k_{\rm r}$ the field and temperature dependent rate coefficients, while a gives the order of the reaction (1 for H^0 and 2 for H_2).

Transport of the hydrogen species away from the interface is assumed to be controlled by conventional diffusion

$$\frac{\partial H}{\partial t} = D\nabla^2 H \ , \qquad (2)$$

with D being the diffusion coefficient. At the boundary we have to consider the influx of the newly created species determined by $a\partial N_{\rm it}/\partial t$. For the calculation of the time

dependent density of interface states, $N_{\rm it}(t)$, (1) and (2) can be solved numerically on an arbitrary geometry. From $N_{\rm it}$ the threshold voltage shift is normally estimated by assuming that all traps are positively charged, an assumption only fulfilled during strong negative bias where the Fermi-level is close to the valence band edge. In addition, any potentially generated oxide charges are often neglected and we obtain

$$\Delta V_{\rm th} = -\frac{\Delta Q_{\rm it}(E_{\rm F}) + \Delta Q_{\rm ot}}{C_{\rm ox}} \approx -\frac{{\rm q}\Delta N_{\rm it}}{C_{\rm ox}} \ . \qquad (3)$$

III. Extended Reaction-Diffusion Model

Although the reaction-diffusion model has been successfully used to describe NBTI, it only qualitatively predicts the relaxation phase. This is demonstrated in Fig. 1 where the NBTI degradation of an 1.3 nm thick oxide during subsequent stress/relaxation cycles is shown in comparison to simulation results obtained from the RD model, assuming H^0 and H_2 kinetics. Good accuracy is only obtained during the first stress phase while in the relaxation phase some saturation is often observed which is not well reproduced. Also, if stress is applied again, the accuracy of the results predicted by the RD model decreases (see also fits to measurements in [3, 15]). This issue is even more important for thicker oxides as used in high-voltage devices where a stronger saturation is often observed [16].

A. Trap-Controlled Transport

As will be shown in the following, the inclusion of hydrogen traps distributed over a wide energy range significantly improves the accuracy of the model. These traps are introduced by replacing the diffusion equation employed in the RD model by dispersive transport models commonly used to describe the motion of the hydrogen species in dielectrics and amorphous materials [17–21]. We base our description on the multiple trapping (MT) model [19, 20, 22] which is compatible to the equations conventionally used in process and device simulation. In the MT model the species $H(\mathbf{x}, t)$ consists of free (conducting) particles $H_{\rm c}(\mathbf{x}, t)$ and particles residing on various trap levels $E_{\rm t}$. The energy density of those trapped particles is given by $\rho(\mathbf{x}, E_{\rm t}, t)$ and the total concentration of trapped hydrogen is calculated as $H_{\rm t}(\mathbf{x}, t) = \int \rho(\mathbf{x}, E_{\rm t}, t)\, {\rm d}E_{\rm t}$. The continuity equation for the total concentration of the species H reads

$$\frac{\partial H_{\rm c}(\mathbf{x}, t)}{\partial t} + \frac{\partial H_{\rm t}(\mathbf{x}, t)}{\partial t} = D\nabla^2 H_{\rm c}(\mathbf{x}, t) \ , \qquad (4)$$

where the flux is only determined by particles in the conduction states. At each trap level a balance equation accounts for the newly trapped particles versus the released ones. The release rate is proportional to the trapped charge on that level, assuming appropriate space in the extended states.

$$\frac{\partial \rho(E_{\rm t})}{\partial t} = c(E_{\rm t}) H_{\rm c}\Big(g(E_{\rm t}) - \rho(E_{\rm t})\Big) - r(E_{\rm t})\rho(E_{\rm t}) \ . \qquad (5)$$

Here, $c(E_{\rm t})$ and $r(E_{\rm t})$ are the energy-dependent capture and release rates, respectively, and $g(E_{\rm t})$ is the trap density-of-states, where commonly an exponential distribution is assumed. In contrast to our previous work [16], we do not assume cracking of H_2 and we also assume that the trapped hydrogen remains uncharged.

Based on simplified solutions for the MT problem [20] NBTI models have been developed [23] by assuming that all hydrogen, free and trapped, can contributed to the reverse interface rate. This boundary conditions results in slopes smaller than those predicted by the RD model but the microscopic justification for such an assumption is still unclear. In contrast, we assume that only free hydrogen can re-passivate a dangling bond, which results in slopes *larger* than predicted by the RD model, consistent with [9].

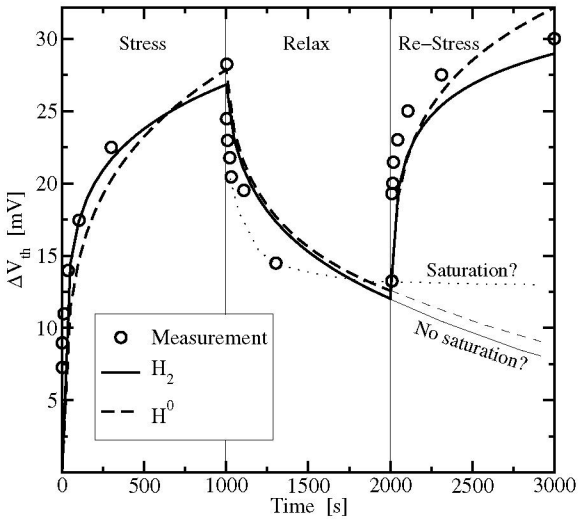

Figure 1: NBTI degradation during subsequent stress/relaxation cycles for a 1.3 nm thick oxide. Although good accuracy can be obtained during the first stress phase, the fit is only qualitative and considerably poorer in the relaxation phase and the second stress phase. Shown are the results of the standard RD model with H^0 and H_2 kinetics together with measurement data from [3].

B. Coupling to the Semiconductor Equations

We solve the hydrogen equations together with the semiconductor device equations, consisting of the Poisson equation and the drift-diffusion equations. To account for the bias dependence of NBTI, the forward rate is conventionally assumed to depend on the electric field. Also, it has been found that holes at the surface are required [3], giving $k_{\rm f} = k_{\rm f0}\, p_{\rm s}/p_{\rm ref}\, \exp(E_{\rm ox}/E_{\rm ref})$. In contrast to standalone implementations of the RD model, the full coupling can now be taken into account, as both $p_{\rm s}$ and $E_{\rm ox}$ are available as solution variables. In addition, the feedback of the created interface charges on $E_{\rm ox}$ is considered which influences the forward rate $k_{\rm f}$ through the exponential dependence on $E_{\rm ox}$. Nevertheless, this feedback effect was found to be small.

More important is the fact that the concentration of amphoteric $P_{\rm b}$ centers, $N_{\rm it}$, needs to be translated to an electrically active interface density-of-states (DOS). In addition to the band-tail states, which are assumed to stay approximately constant during stress, the newly created $P_{\rm b}$ centers introduce two

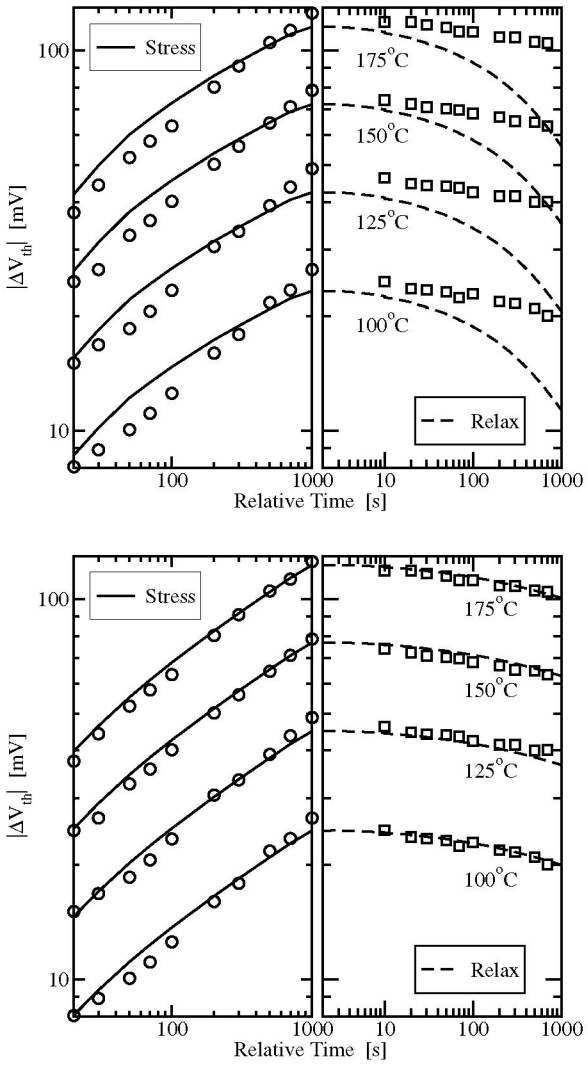

Figure 2: Comparison of measurement results with a calibrated RD simulation (top) and the extended RD model (bottom). The discrepancy observed in the relaxation phase of the RD model is removed by the inclusion of traps. Also, the temperature-independent slope of $n = 0.31$ during the stress phase is better matched with the trap-controlled transport model.

dynamically changing distinct peaks in the Si bandgap [24]. We model the change in the DOS as

$$\Delta g(E_\mathrm{t}) = N_\mathrm{it}\Big(g_\mathrm{P}(E_\mathrm{t}, E_\mathrm{P1}, \sigma_1) + g_\mathrm{P}(E_\mathrm{t}, E_\mathrm{P2}, \sigma_2)\Big) \quad (6)$$

with

$$g_\mathrm{P}(E_\mathrm{t}, E_\mathrm{P}, \sigma) = \frac{\gamma}{\sigma}\frac{\exp\Big(\frac{E_\mathrm{P}-E_\mathrm{t}}{\sigma}\Big)}{\Big(1+\exp\Big(\frac{E_\mathrm{P}-E_\mathrm{t}}{\sigma}\Big)\Big)^2}\,. \quad (7)$$

and $E_\mathrm{P1} - E_\mathrm{v} = 0.235\,\mathrm{eV}$, $E_\mathrm{P2} - E_\mathrm{v} = 0.85\,\mathrm{eV}$, and $\sigma_1 = \sigma_2 = 0.08\,\mathrm{eV}^2$ [25]. The correction factor γ is introduced to guarantee that the integral over one peak in the band-gap equals N_it, in order to allow for comparison with models that assume all interface states to be positively charged. Here, the peak around E_P1 is donor-like, that is, can be positively charged or neutral, while peak E_P2 is acceptor-like and not

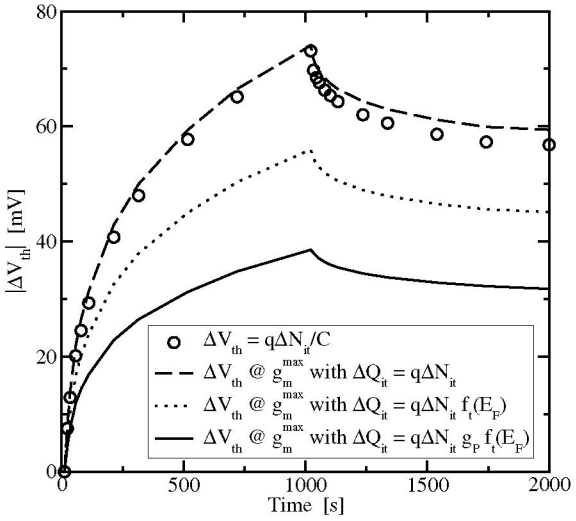

Figure 3: Calculated threshold voltage shift assuming three different charge distributions at the interface in comparison to the V_th shift predicted by (3). V_th is extracted at maximum g_m from the simulated $I_\mathrm{D}V_\mathrm{G}$ curves assuming a) that all states are positively charged, b) that the DOS is constant inside the bandgap, c) that the DOS follows (6).

normally charged during operation of pMOS devices. The DOS (6) is used in a surface SRH recombination mechanism to properly account for the dynamically changing trap occupancy, in particular during measurement cycles.

IV. Example and Discussion

To evaluate the model we consider a high-voltage pMOS structure with a gate oxide thickness of 48 nm which was stressed at $T_\mathrm{L} = 100, 125, 150,$ and $175\,^\circ\mathrm{C}$ by applying $V_\mathrm{G} = -25\,\mathrm{V}$. The degradation was monitored by interruption the stress for 2 s at regular intervals where a full $I_\mathrm{D}V_\mathrm{G}$ curve was measured from which V_th was determined at maximum g_m. Note that for these devices delay-free measurements have not yet been shown to be feasible due to the large gate overdrive and the resulting current levels. Therefore, conventional stress-measurement-stress techniques have been used. In that context it is important to consider the exact measurement setup during the simulation by applying the same time-dependent bias conditions and extracting V_th in the same way as in the measurement. Calibration results of the RD model are shown in Fig. 2 (top), where particularly in the relaxation phase a strong discrepancy is observed. Application of the extended RD model, however, leads to a significant improvement as shown in Fig. 2 (bottom). Best results were obtained by adding a deep trap level at $\approx 0.5\,\mathrm{eV}$.

Another important issue is the influence of the partial occupancy of the newly created P_b centers on the extracted V_th shift and the subthreshold slope. Although during stress the assumption $Q_\mathrm{it} \approx \mathrm{q}N_\mathrm{it}$ is accurate, during the $I_\mathrm{D}V_\mathrm{G}$ sweep the trap occupancy changes as the Fermi-level moves closer to mid-band. The influence on the extracted V_th shift is shown in Fig. 3 for three different models for Q_it: Assuming that $Q_\mathrm{it} \approx \mathrm{q}N_\mathrm{it}$ irrespectively of the Fermi-level basically reproduces the result of (3) and does not change the subthreshold slope.

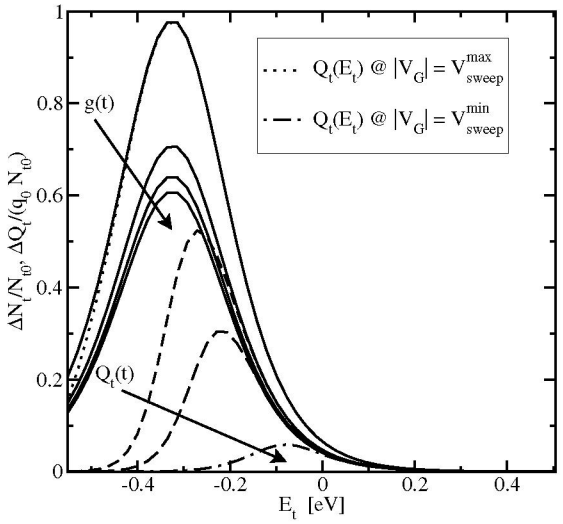

Figure 4: Relative occupancy of the donor-like interface trap levels during a measurement cycle, where $V_{\rm G}$ goes from $V_{\rm G}^{\rm max}$ to $V_{\rm G}^{\rm min}$. As the surface potential comes closer to the mid-gap position, the trap occupancy is considerably reduced, while $N_{\rm it}$ degradation recovers only slowly.

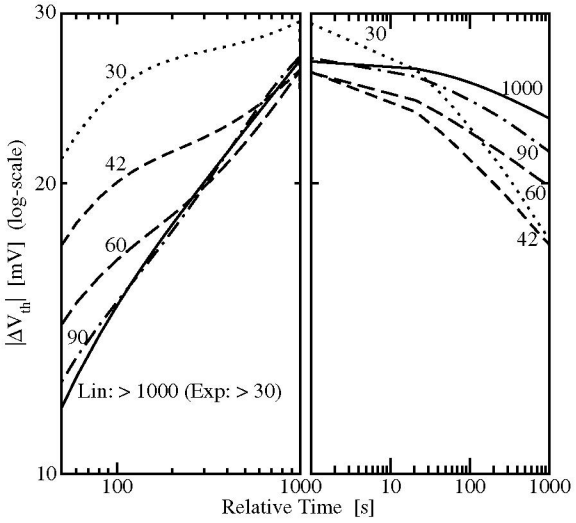

Figure 5: A strong sensitivity of the simulation result on the number of mesh points is observed. For an equidistant mesh, a large number of mesh points is required, while, in this example, for an exponentially increasing mesh spacing 30 mesh points are sufficient.

For a constant DOS and with proper consideration of the trap occupancy, the extracted $V_{\rm th}$ is lower, as not all trap levels are occupied anymore as shown in Fig. 4. Finally, when accounting for the peaks introduced by the $P_{\rm b}$ centers, the trap occupancy becomes even smaller during $V_{\rm th}$ extraction since the peak position used in (6) is relatively close to the valence band, resulting in an even smaller shift of $V_{\rm th}$. It is important to note that the constant and peak DOS as used in these simulations are only extreme cases of the more realistic scenario where also fixed interface charges are created [2], thereby reducing the bias dependence of $Q_{\rm it}(E_{\rm F})$. Also note, that on-the-fly measurements which monitor the change in $I_{\rm D}$ without reducing the stress voltage, are not susceptible to that issue.

Finally, a strong sensitivity of the simulation result on the number of mesh points was observed. This is because hydrogen is released close to the Si/SiO_2 interface where it can diffuse back to repassivate an interface state. Therefore, when an equidistant mesh is used in the whole oxide, a large number of mesh points ($N > 1000$) is required to properly resolve the critical near-interface region. The problem can be resolved by using an exponentially growing mesh spacing and $N = 30$ was found to be sufficient in that particular example.

V. Conclusions

We have proposed and discussed a coupled solution of an extended reaction-diffusion model with the semiconductor device equations. It was found that trap-controlled transport allowed to considerably improve the accuracy in the stress and the relaxation phase. Furthermore, it was found that the density-of-states of the electrically active interface states is crucial as it considerably influences the predicted $V_{\rm th}$ shifts.

References

[1] A. Goetzberger and H. Nigh, Proc.IEEE **54**, 1454 (1966).
[2] D. Schroder and J. Babcock, J.Appl.Phys. **94**, 1 (2003).
[3] M. Alam and S. Mahapatra, Microelectr.Reliab. **45**, 71 (2005).
[4] S. Mahapatra *et al.*, Microelectr.Eng. **80**, 114 (2005).
[5] V. Huard *et al.*, Microelectr.Reliab. **46**, 1 (2006).
[6] J. Stathis and S. Zafar, Microelectr.Reliab. **46**, 270 (2006).
[7] A. Krishnan *et al.*, Appl.Phys.Lett. **88**, 1 (2006).
[8] J. Yang *et al.*, Appl.Phys.Lett. **88**, 1 (2006).
[9] S. Chakravarthi *et al.*, in *Proc. IRPS* (2004), pp. 273–282.
[10] H. Kufluoglu and M. Alam, IEEE Trans.Electr.Dev. **53**, 1120 (2006).
[11] K. Jeppson and C. Svensson, J.Appl.Phys. **48**, 2004 (1977).
[12] P. Lenahan and J. Conley Jr., J.Vac.Sci.Technol.B **16**, 2134 (1998).
[13] J. Campbell *et al.*, Appl.Phys.Lett. **87**, 1 (2005).
[14] D. Varghese *et al.*, in *Proc. IEDM* (2005), pp. 1–4.
[15] A. Krishnan *et al.*, in *Proc. IEDM* (2003), pp. 1–4.
[16] R. Entner *et al.*, in *Proc. WoDiM* (2006), pp. 96–97.
[17] E. Montroll and H. Scher, J.Stat.Phys **9**, 101 (1973).
[18] H. Scher and E. Montroll, Phys.Rev.B **12**, 2455 (1975).
[19] J. Noolandi, Phys.Rev.B **16**, 4466 (1977).
[20] V. Arkhipov and A. Rudenko, Philos.Mag.B **45**, 189 (1982).
[21] D. Brown and N. Saks, J.Appl.Phys. **70**, 3734 (1991).
[22] J. Orenstein *et al.*, Philos.Mag.B **46**, 23 (1982).
[23] B. Kaczer *et al.*, in *Proc. IRPS* (2005), pp. 381–387.
[24] A. Haggag *et al.*, in *Proc. IRPS* (2001), pp. 271–279.
[25] L.-A. Ragnarsson and P. Lundgren, J.Appl.Phys. **88**, 938 (2000).