Modeling Bias Temperature Instability During Stress and Recovery

Tibor Grasser,* Wolfgang Goes,* and Ben Kaczer†

*Christian Doppler Laboratory for TCAD at the Institute for Microelectronics, TU Wien
Gußhausstraße 27–29/E360, A–1040 Wien, Austria
Email: {grasser|goes}@iue.tuwien.ac.at
† IMEC, Kapeldreef 75, B–3001 Leuven, Belgium
Email: kaczer@imec.be

***Abstract*—Bias temperature instability has attracted a lot of attention as a dominant degradation mechanism in modern MOS transistors. Despite considerable effort, the exact physics behind this mechanisms are still controversial. We discuss some numerical aspects of our recently presented model which is capable of reproducing the main features of the phenomenon. Furthermore, we demonstrate how the model can be applied to understand variations in nominally identically stressed devices which have become important in the small area limit.**

I. Introduction

Judging from the number of recent publications at leading reliability conferences, bias temperature instability (BTI) is one of the most researched reliability issues in modern MOS transistors. Nevertheless, even after four decades of research, it is still a highly puzzling phenomenon which has so far eluded our complete understanding [1–3]. BTI is observed when a large negative (or positive) voltage is applied to the gate of a MOSFET which causes a shift of the threshold voltage and other crucial transistor parameters. The degradation is considerably accelerated at elevated temperatures. Particularly intriguing are the complex degradation/recovery patterns which are observed when the gate bias is modulated [4–6]. Understanding this behavior is mandatory for any estimation of device degradation in a circuit setting.

An important aspect of the degradation is that it seems to consist of a recoverable and a permanent component. While the recovery which can cover a large range of time scales, e.g. from $1\,\mu$s up to $1\,$Ms [7], has been unequivocally observed by many groups, it is still under debate whether the permanent component is completely permanent [1, 3]. Also, there is a considerable controversy whether and how to map each component on physical processes like hole-trapping and interface state generation.

Quite contrary to recent experimental observations, most modeling approaches published so far have focused on constant gate bias stress and are remarkably oblivious to any recovery of the degradation which sets in as soon as the stress is removed. In particular, the popular reaction-diffusion (RD) theory [8] has been shown to be inadequate [9] for the description of BTI relaxation, see Fig. 1.

Most intriguingly, degradation data taken at different stress temperatures and voltages appear to follow a universal pattern [3]: at any stress and relaxation time the $V_{\rm th}$ shift can be expressed as $\Delta V_{\rm th}(T, V_{\rm s}, t_{\rm s}, t_{\rm r}) \approx s(T, V_{\rm s})\Delta V_{\rm th}^0(t_{\rm s}, t_{\rm r})$, with the universal function $\Delta V_{\rm th}^0$ containing the dependence on the stress and relaxation times $t_{\rm s}$ and $t_{\rm r}$ an a prefactor $s(T, V_{\rm s})$ containing the temperature T and voltage $V_{\rm s}$ dependence. This scalability strongly suggests that BTI is either the result of a single process or of tightly coupled processes. In particular, the previously suggested mere superposition of independent hole-trapping and interface state generation mechanisms [1] is incompatible to this observation.

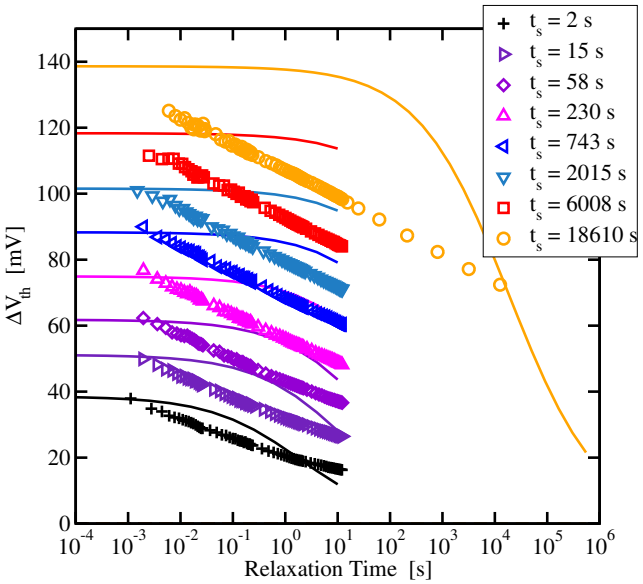

Fig. 1: Relaxation of $\Delta V_{\rm th}$ after 8 different stress times as predicted by the reaction-diffusion model in comparison to measurement data. The poor correlation of the RD model with the data is striking. In particular, the measurement data behave like $\log(t_{\rm r})$ over many decades in time while the RD model predicts no initial recovery followed by a fast transition to zero around $t_{\rm relax} = t_{\rm stress}$.

II. The Triple-Well Model

In a first attempt to construct a model that is able to reproduce both a recoverable on top of a permanent component while being compatible with the scalability property we have recently suggested a triple-well energy model [10], see Fig. 2. The model is based on the dissociation of Si–H bonds inside the nitrided SiO_2 insulator. As a result, positively charged defects are created which lead to a shift of $\Delta V_{\rm th}$ and other

978-1-4244-1753-7/08/$25.00 ©2008 IEEE

device parameters. The dissociation is assumed to proceed predominantly in two steps which are modeled via their energetic configuration, where the lowest energetic position corresponds to the silicon hydrogen bond (cf. Fig. 2). Upon application of the electric field, H is moved to well 2 (transition from V_1 to V_3 over the barrier V_2). Once in well 2, electrically active states are assumed to appear inside the silicon bandgap [11] and capture of a hole leads to a further reduction of the barrier and complete dissociation of the hydrogen atom (transition from V_3 to V_5 over the barrier V_4). This second step has a larger barrier and consequently leads to a more permanent deviation from the equilibrium configuration. After removal of the stress, particles residing in well 2 move back rather quickly to the equilibrium configuration while particles in well 3 require considerably larger times.

The rate equations describing this dynamic process together with the transition rates are given as

$$\frac{\partial f_1}{\partial t} = -f_1 k_{13} + f_3 k_{31} \ , \tag{1a}$$

$$\frac{\partial f_3}{\partial t} = +f_1 k_{13} - f_3 k_{31} - f_3 k_{35} + f_5 k_{53} \ , \tag{1b}$$

$$\frac{\partial f_5}{\partial t} = +f_3 k_{35} - f_5 k_{53} \ , \tag{1c}$$

$$f_1 + f_3 + f_5 = 1 \tag{1d}$$

with the reaction rates

$$\begin{aligned} k_{13} &= \nu \exp(-\beta(V_2 - V_1 - \Delta_2)) \ , \\ k_{31} &= \nu \exp(-\beta(V_2 - V_3 + \Delta_2)) \ , \\ k_{35} &= \nu \exp(-\beta(V_4 - V_3 - \Delta_4)) \ , \\ k_{53} &= \nu \exp(-\beta(V_4 - V_5 + \Delta_4)) \ . \end{aligned}$$

The probability of the hydrogen atom being in well one is given by f_1, the probability of it being in well 2 and 3 f_3 and f_5, respectively, while having a defect is the probability of the particle not being in well 1, thus $1 - f_1$. Furthermore, we use $1/\beta = \mathrm{k_B}T$ and an attempt frequency $\nu \approx 10^{13}\,\mathrm{s}^{-1}$. During stress, Δ_2 and Δ_4 describe the modification of the barrier heights [12]. The equation system describing a single reaction path is linear and easy to solve analytically.

Since the oxide is amorphous, the energies describing the reaction path are assumed to be random variables [13], for instance given as $\mathbf{X} = (V_1, V_2, V_3, V_4, V_5)$. Other random variables determining the properties of a possible reaction path may be easily envisaged, for instance its orientation with respect to the applied electric field, ϑ, and its distance from the interface, x. The probability density of the reaction paths is given by $g(\mathbf{X})$ and consequently we obtain for the shift of the threshold voltage

$$\Delta V_{\mathrm{th}}(t) = -\frac{\mathrm{q}}{C_{\mathrm{ox}}} \int \mathrm{d}\mathbf{X}\ g(\mathbf{X}) \Big(1 - \Delta f_1(t, \mathbf{X})\Big) \Big(1 - \frac{x}{t_{\mathrm{ox}}}\Big) \ . \tag{2}$$

In the following we assume the defects to be close to the $\mathrm{Si/SiO_2}$ interface ($x \approx 0$) and $g(\mathbf{X})$ to be given by independent Gaussian distributions.

The triple-well model is evaluated against measurement data taken on $t_{\mathrm{ox}} = 1.4\,\mathrm{nm}$ oxynitrided MOSFETs. Excellent agreement has been found during both stress and recovery, after positive and negative bias stress, during mixed positive and negative stress, and during duty-factor dependent stresses [10], see Figs. 3 and 4 for selected examples. Here we further explore some features of the model along with its numerical properties.

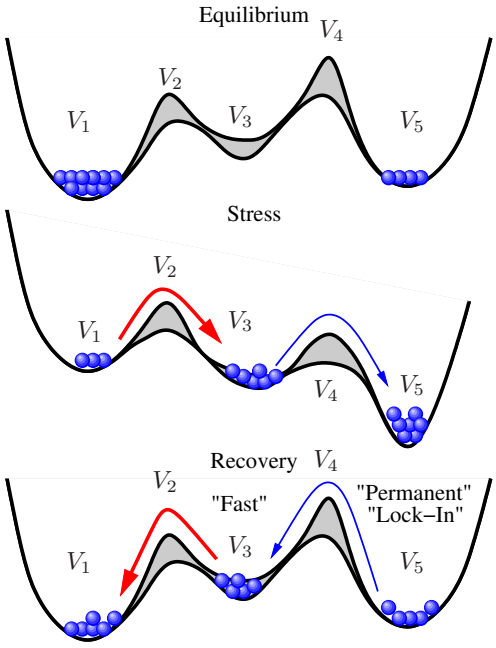

Fig. 2: The energy levels involved in the triple-well model. The second well is energetically higher than the first and the third well and forms an intermediate configuration. Transitions from the second well back to the first well are fast, while the third well represents the permanent component/lock-in.

III. Numerical Solution Procedure

Depending on the number of random variables, direct discretization of (2) leads to a very large number of equation systems of the form (1). For instance, assuming that every random variable is discretized using 20 points, we have to solve 8.000 equations for $\mathbf{X} = (V_2, V_3, V_4)$ but already 160.000 for $\mathbf{X} = (V_2, V_3, V_4, V_5)$ and more than 3 millions for 5 random variables. In addition, direct discretization of a large number of random variables leads to undesired correlations between the reaction paths. As an example, a too crude discretization of V_2, the barrier that dominates the initial degradation behavior, gives unwanted 'steps' in the initial simulation result, despite the large total number of reaction paths (see Fig. 5).

We therefore proceed as follows: Rather than looking for a rigorous solution of (2), the correct limit for an infinitely large (nominal) device, we solve (2) using a statistical approach under consideration of the real device dimensions. Assuming an effective maximum defect density of $10^{13}\,\mathrm{cm}^{-2}$, we obtain for our devices with $A = W \times L = 100\,\mathrm{nm} \times 1\,\mu\mathrm{m}$ a total number of possible defects equal to $N \approx 10^4$. The dissociation path associated with each defect will be slightly different and is determined independently from the other paths. Summation of the contributions from the individual reaction

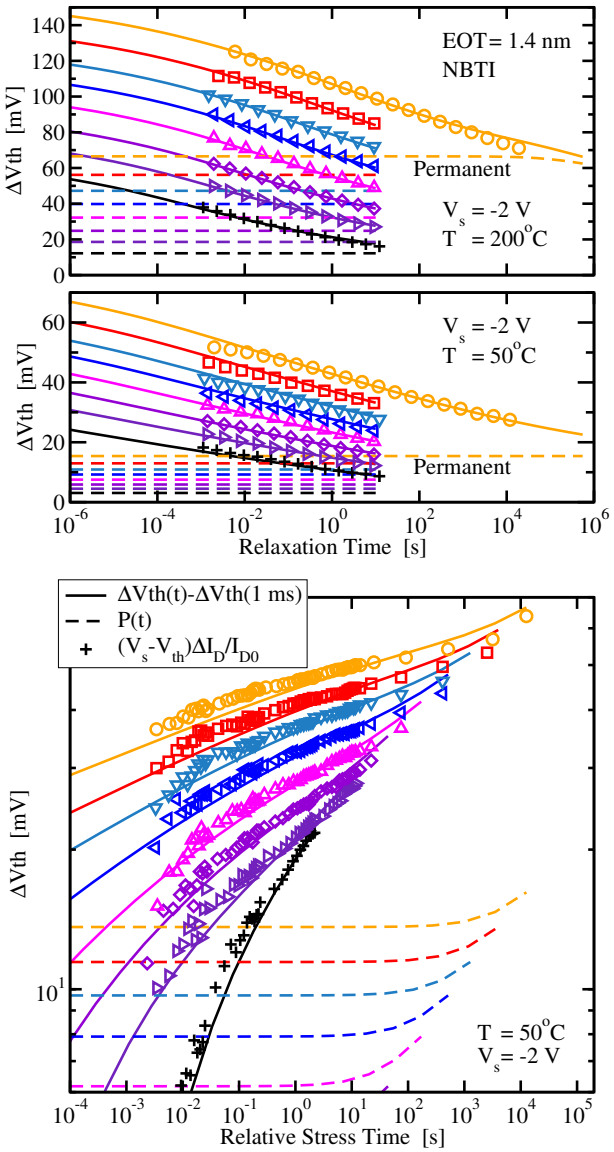

Fig. 3: Evaluation of the triple-well model (solid lines: total degradation, dashed lines: 'permanent' component in well 3) against the same data as in Fig. 1. Note that all data were recorded in a single measurement sequence which was interrupted 8 times to record the relaxation data (same stress times as in Fig. 1). For the model evaluation the measurement sequence was simulated using fixed parameters in a single simulation run, just like the real measurement. **Top**: Excellent accuracy is obtained for two temperatures during the relaxation phase. **Bottom**: The threshold-voltage shift predicted by the triple-well model is also in excellent agreement during each stress phase (same symbols and colors for the stress curve preceding the relaxation phase in the left figure). Also, the initial delay in the first measurement was taken into account by subtracting $\Delta V_{\rm th}(1\,{\rm ms})$ from the simulation result.

paths following

$$\Delta V_{\rm th}(t) = -\frac{\rm q}{C_{\rm ox}} \sum_{i=1}^{N} \Big(1 - \Delta f_{i,1}(t)\Big)\Big(1 - \frac{x_i}{t_{\rm ox}}\Big) \quad . \qquad (3)$$

then gives the total device degradation. Depending on the seed of the random number generator used to create the random configurations, a slightly different $\Delta V_{\rm th}$ under the same stress conditions is obtained from (3). However, for sufficiently large N, (3) gives an excellent approximation of (2).

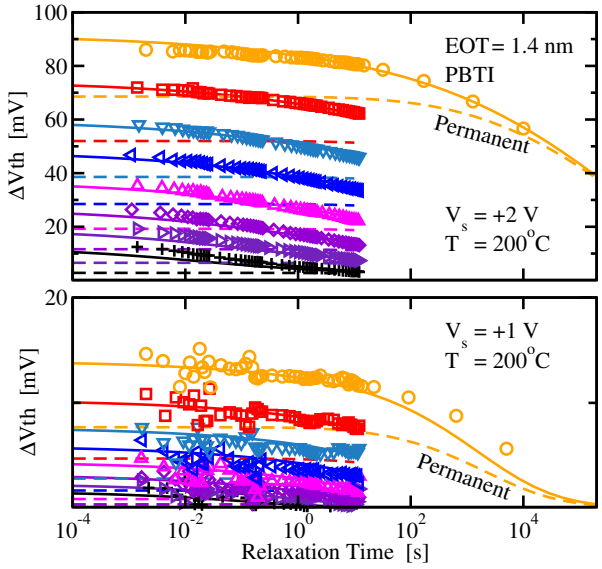

Fig. 4: The triple-well model can also be used to describe PBTI stress. During positive bias stress, less permanent degradation is created. Note that PBTI stress also creates positive charge and thus a negative $\Delta V_{\rm th}$, just like NBTI.

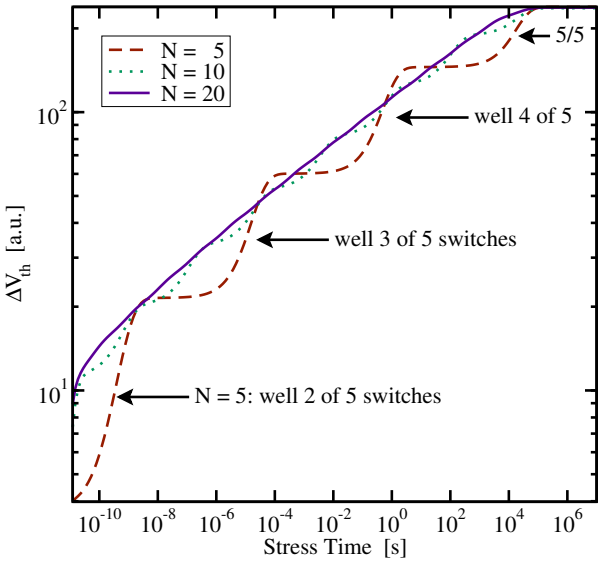

Fig. 5: Influence of the number of discretization points on the initial response of the triple-well model. When a small number is used, the individual transitions of each well from its equilibrium configuration to the final stress equilibrium configuration becomes clearly visible. A large number of discretization points, on the other hand, leads to an excessive number of reaction paths.

IV. Small Area Devices

For small area devices, however, device-to-device variations do exist and a deviation from the nominal behavior is observed. These stress-induced intrinsic fluctuations are a real concern for analog-matched pMOSFETs and SRAM stability [14]. Being caused by discrete charges homogeneously distributed underneath the gate [15], their variance scales with $\sigma^2(\Delta V_{\rm th}) \sim 1/A \sim 1/N$, just like random-dopant-induced fluctuations [15, 16]. This is consistent with the nominal behavior that is obtained for a reasonably large N.

In our model device-to-device variations can be naturally