A Two-Stage Model for Negative Bias Temperature Instability

T. Grasser†, B. Kaczer‡, W. Goes*, Th. Aichinger†, Ph. Hehenberger*, and M. Nelhiebel‡

* Christian Doppler Laboratory for TCAD at the † Institute for Microelectronics, TU Wien, Austria
‡ IMEC, Leuven, Belgium ‡ KAI, Villach, Austria ‡ Infineon Technologies, Villach, Austria

Abstract—Based on the established properties of the most commonly observed defect in amorphous oxides, the E’ center, we suggest a coupled two-stage model to explain the negative bias temperature instability. We show that a full model that includes the creation of E’ centers from their neutral oxygen vacancy precursors and their ability to be repeatedly charged and discharged prior to total annealing is required to describe the first stage of degradation. In the second stage a positively charged E’ center can trigger the depassivation of P3 centers at the Si/SiO2 interface or K0 centers in oxynitrides to create an unpassivated silicon dangling bond. We evaluate the new model to experimental data obtained from three vastly different technologies (thick SiO2, SiON, and HK) and obtain very promising results.

I. INTRODUCTION

Many recent publications dealing with the negative bias temperature instability (NBTI) have suggested the existence of a recoverable component observed on top of a slowly recovering or even permanent component [1–4]. Often, the recoverable component is attributed to hole trapping while the permanent component is explained by the creation of interface states [1]. We have recently pointed out a serious problem with the interpretation that two independent components result in the overall degradation observed during NBT stress [4, 5]. This is because these two components should have a different voltage and temperature acceleration, allowing for their separation by the application of a suitably chosen combination of stress temperatures and voltages. Quite to the contrary, however, we have observed [4–7] that NBTI data at various stress and relaxation times, broad ranges of stress voltages and temperatures, and we have broad range of technologies (ultra-thin SiON, high-k, ultra-thick SiO2) can often be made to overlap via multiplication by a suitably chosen scaling factor. This very broad scalability cannot be explained by any established model [4] but implies that NBTI is either due to a single mechanism (which then would have to be able to explain both the recoverable and the permanent contributions) or due to two tightly coupled mechanisms. We suggest a model that captures this behavior during both stress and recovery and explain why previous modeling attempts fail in doing so.

II. EXPERIMENTAL OBSERVATIONS

We characterized SiON pMOSFETs (EOT = 1.4nm) using the extended MSM scheme (eMSM) [3, 8], which acquires data in alternating on-the-fly (OTF) and relaxation sequences using a wide range of stress voltages (−0.6V . . . −2.0V) and stress temperatures (25°C . . . 200°C). The recorded OTF degradation in Ith was converted to ΔVth using the simple expression ΔVth ≈ (Vth − Vth0) / Ith0 / Ith0 [9, 10]. It is now understood that ΔVth contains information regarding the time dynamics of the initial drain current and the error in the initial drain current. We have suggested [13, 14], but are still open to rigorous justification. During model development, the impact of Ith can be easily accounted for by subtracting the simulated ΔVth(t0) from the overall simulation result. The impact of the mobility variation, however, remains unclear at the moment.

Fig. 1: Degradation of the drain current collected during consecutive stress sequences. The stress temperature was varied at a fixed stress voltage of −2V. The unscaled data initially follows a logarithmic time dependence while the long-time data may be approximated by a power-law with the exponent n = 0.11. The slope of the initial log shows clear temperature activation while the power-law exponent is roughly temperature-independent. Multiplication of each data set with a constant value results in a nearly perfect overlap. The scaling factors are independent of stress time, indicating that the initial and the long-time power-law behavior are due to a related process.

Nonetheless, we will take the uncorrected ΔVth as an indicator for the overall degradation and assume that ΔVth contains the correct information regarding the time dynamics but may potentially be affected by an unknown error in amplitude [13].

As has been observed previously [1, 4, 6, 15, 16], for short stress times (t ≤ 1 s, depending on the stress condition) the initial degradation phase is well described by a logarithmic time-dependence (see Fig. 1 for a typical example),

\[ \Delta V_{th}(t_0) = B_s(F,T) \log_{10}(t_0/t_i), \]  

with \( t_0 \) being the delay of the first measurement point. The pre-factor during stress, \( B_s \), gives the increase in \( \Delta V_{th} \) in volts per decade in time and depends on the stress field \( F \) and temperature \( T \).

The recovery phase, on the other hand, may be fit by [1, 17–19]

\[ \Delta V_{th}(t, t_i) = B_r(F,T) \log_{10}(1 + t_i/t_i) + P(t_i), \]  

where \( P \) is a roughly permanent contribution depending on the stress time only and is possibly due to interface states [17]. Again, the pre-factor \( B_r \) gives the recovery rate in volts per decade in time.

Before developing our model, we briefly summarize some key experimental observations obtained by the eMSM technique, which provide important insights regarding the dynamics of NBTI. In particular, they can be used to effectively rule out a number of alternative models as will be shown later.
A. Scalability

The data recorded using the eMSM scheme can often be made to overlap by multiplication with a suitably chosen scaling factor. While the initial behavior up to about 1 s is well approximated by a logarithmic time dependence and the long-term data approximately follows a power-law, the same scaling factor can be used in both regimes, see Fig. 1. A similar scalability is observed during recovery, see Fig. 2 which, to first-order, follows a logarithmic time dependence. We remark that there is an excellent correlation between the scaling factors required during stress and recovery [6].

B. Bias and Temperature Dependence of Stress

Following the previously suggested idea that the initial degradation is dominated by hole trapping while the long-term degradation is due to interface state creation, we focus on the initial degradation phase to obtain some more information on this alleged ‘hole trapping’ component. We thus subjected devices to short stresses of 2 s and let them recover for about 3000 s, see Fig. 3. For what is typically considered NBTI stress in this oxide thickness range (Vstress ≤ −1 V), the prefactor Bt can be roughly approximated as Bt(F,T) ≈ Bt,0 Tn F2, with Θ ≈ 2. We remark that the power-law temperature dependence is also well approximated by an Arrhenius law with EA ≈ 70 meV.

C. Asymmetry Between Stress and Relaxation

It has been long understood that recovery takes substantially longer than the time used to build up the degradation. This is shown in Fig. 4 for a typical stress temperature of 125°C. Using the extracted pre-factors Bt (fit to a log in the range 1 ms...1 s) and Bt (fit to a log in the range 1 ms...100 ms) we observe that the ratio Bt/Bt is about 2.5, independent of temperature and voltage. For example, if one observes during degradation a rate of 10 mV/dec, recovery will proceed with only about 4 mV/dec. The exact value of Bt/Bt depends on the mobility error in Bt but this has no bearing on the fact that stress and recovery are asymmetric. Although this asymmetry may look quite innocent at a first glance, it turns out to be a considerable challenge for any modeling attempt.

D. Bias Dependence of Recovery

Recovery has been shown to depend on the bias voltage applied during recovery [1, 5, 8, 19]. In particular, application of positive bias accelerates recovery. This has been explained by the bias-dependence of H+ drift [8] or the bias-dependence of hole-detrapping via valence band and interface states [1]. However, just as the asymmetry, this is also challenging to reproduce correctly by a quantitative model.

III. Previous Modeling Attempts

Although it has been widely acknowledged that both interface states and trapped holes can contribute to NBTI, the exact details on how this should occur are highly controversial. The vastly different microscopic explanations that have been suggested will be quickly summarized in the following.
A. Interface States Only

In the simplest case NBTI would be due to a single mechanism only. One prominent example is the reaction-diffusion (RD) theory which claims that only interface states are responsible. However, it has been clearly shown that the RD theory and the various extensions proposed over the years can only explain constant bias stresses while they fail to explain the dynamics (recovery, bias dependence of recovery, duty-factor dependence, etc.) [2, 4, 15]. Furthermore, published data on the recovery of charge-pumping (CP) signals seem to indicate that interface states show only marginal recovery compared to the recovery of $\Delta V_{th}$ [1, 20].

Data acquired by the recently suggested on-the-fly charge-pumping technique (OFIT), which tries to minimize the measurement-induced recovery by using the CP base-level as a stress voltage [21], suggested that also interface states can show fast recovery. Consequently, a single mechanism model based on interface states only appeared feasible and we have recently suggested such a model [5]. The model is derived using a suitable generalization of dispersive bond breaking already used previously for the creation of interface states [1, 22]. Although the resulting triple-well model can reproduce complicated stress/relaxation sequences with very good accuracy, it requires a large and (as it now seems) nonphysical variance in the hydrogen binding energies in order to reproduce data recorded in a larger temperature and voltage range such as used in this study. Furthermore, our theoretical and experimental study of the OFIT technique suggests that fast recovery of interface states may be an artifact of the method and requires correction [23]. We conclude from this that the microscopic assumptions underlyng the triple-well model are likely not correct. Nonetheless, the mathematical structure of the model can be retained in the following.

B. Hole Trapping Models

Hole trapping is often modeled using elastic tunneling into pre-existing traps located at various distances away from the interface. Depending on the distance, an exponentially increasing time constant is obtained, which, at least in thicker oxides, could explain the large spread of time constants observed in NBTI. However, elastic hole trapping is to first-order temperature independent and linearly dependent on the stress field [24]. Furthermore, the model predicts $B_s \sim B_r$, that is, symmetric degradation and recovery behavior. As such, this is incompatible with our data.

C. Combination of Hole Trapping and Interface State Creation

Two variants of combined hole trapping and interface state creation are commonly used: (i) initial hole trapping which quickly saturates (within about 1 s) and long term degradation dominated by interface state creation according to the RD theory [25]. We rule out this variant based on the various known shortcomings of RD theory [2, 4, 15]. (ii) Trapped holes are responsible for the recoverable part of the degradation while interface states form a permanent contribution [1]. We also rule out this variant due to the shortcomings of elastic tunneling and the lack of scalability [4]. (iii) Alternatively, it has been speculated that holes can be trapped into newly created defects [26, 27]. Although the details behind the responsible process have only been schematically outlined, they appear broadly consistent with our own observations and the detailed physical model suggested in the following.

IV. HOLE TRAPPING PROCESS

Since the corrected OFIT data do not indicate fast recovery of interface states [23] and elastic hole tunneling also cannot explain the data, we have to look for an alternative explanation. Of particular interest are hole trapping models that have been applied since the 1970s in attempts to understand 1/f-noise and thermally stimulated currents at semiconductor surfaces [28, 29]. Just like the triple-well model, these models are also based on a dispersion of activation energies but require very large variances in order to reproduce for instance the 1/f behavior. In these models it is assumed that holes can be captured via a (thermally activated) multiphonon emission (MPE) process into deep near-interfacial states/border traps [24, 29–31], for instance into oxygen vacancies ($E'$ centers) [29]. The MPE process differs from the conventionally invoked elastic tunneling process, notably due to its temperature activation and the larger time constants resulting therefrom [30].

Unfortunately, a trapping model based on the MPE process would still have a linear field dependence in contradiction to our data.
In order to resolve this issue, we have to recall that the MPE mechanism is derived under the assumption of negligible electric fields. This assumption is definitely violated in the case of NBTI. An extension of MPE to the large electric field case has already been developed for the emission of particles from deep traps and has become known as multiphonon-field-assisted tunneling (MPFAT) [32,33]. The signatures of this mechanism are its $\exp(F^2/F_0^2)$ field dependence (note that only the logarithm of the enhancement factor enters $B$) and a considerable temperature activation (from the MPE process). The MPFAT process is schematically illustrated in Fig. 5: a hole can either be in the valence band or in a trapped state. These two states are represented by the two solid parabola which give the total energy of the system. The vibrational modes can be approximated using a simple oscillator model and at the intersection point a transition can occur. The intersection point determining the barrier $\Delta E_B$ is dramatically lowered by the application of an electric field, resulting in an enhancement of $\exp(F^2/F_0^2)$.

V. PROPERTIES OF THE $E'$ CENTER

In order to develop an accurate microscopic model for hole trapping in the context of NBTI, we summarize the most important features collected in a long line of studies on oxide defects providing a solid basis for our NBTI model.

- The most likely microscopic candidate for the ’trapped hole’ are defects from the $E'$ center family, most notably the $E'_c$ center [34, 35]. An $E'_c$ center is thought to be created when a hole is trapped in the precursor structure, which is commonly assumed to be a neutral oxygen vacancy (a Si–Si dimer inside the oxide).
- The energy level for hole trapping is roughly located at about 1 eV below the Si valence band [34], see Fig. 6.
- Once the silicon bond is broken, the distance between the two silicon atoms increases into a new equilibrium position, which requires a large-range structural relaxation of the surrounding lattice (10 Å [36]), and an $E'_c$ center is obtained. The $E'_c$ is visible in ESR (electron spin resonance) when positively charged, that is, right after hole capture.
- An important peculiarity about the $E'_c$ center is that it can be repeatedly charged and discharged. The corresponding energy levels lie within the silicon bandgap [34]. The idea behind this cyclability is that once the hole is emitted (that is, an electron is captured), the bond between the two silicon atoms does not fully reform but the defect remains in a dipole state which can easily lose an electron again. The fact that the $E'$ center can act as a switching trap has been suggested by Lelis et al. [37] based on electrical measurements which then suggested the Harry-Diamond-Laboratories (HDL) model. This was later confirmed by ESR studies [38] and theoretical calculations [39], see Fig. 7 for a schematic representation. In [40] it has been suggested that in order to create a stable $E'$ configuration from an oxygen vacancy, the doubly positive configuration could be important.

- Only after having been in its electrically neutral state for a while, the structure relaxes again to the initial dimer configuration and the defect is completely healed.
- The $E'_c$ is often considered a donor-like defect [34], that is, it is either neutral (ESR inactive) or positively charged (ESR active).
- The oxygen vacancy can also act as an electron trap, with a trap level close to the silicon conduction band [39,40].

- Due to the amorphous nature of the interfacial layer, a considerable broadening of the energy-levels is to be expected. For instance, theoretical calculations give a spread of about 1.5 eV for the precursor level and a Si–Si bond-length variation from 2.3 Å to 2.7 Å [36].

The above summary bears some important and interesting consequences regarding our understanding of NBTI:

- So far, hole trapping has been mainly considered as being into pre-existing traps which rapidly fill but are not related to the actual NBTI mechanism, a somewhat parasitic component which has to be removed to get to the heart of NBTI [10].
- Using CV measurements, NBTI has been shown to be due to donor-like defects [10,41,42]. Models based on the RD theory claim that NBTI is dominantly due to the creation of interface states, which are $P_h$ centers. However, $P_h$ centers are amphoteric, that is, donor-like only in the lower half of the silicon band-gap but acceptor-like in the upper half [43]. The $E'_c$ center, on the other-hand, is a donor-like defect [34].
- NBTI recovery is strongly bias-dependent, in particular when the gate voltage is moved from inversion into accumulation [8]. This is intuitively consistent with carrier trapping [27].
- Describing hole trapping via the known properties of the $E'$ center promotes the positive oxide charge component from a purely parasitic component to the central contributor to NBTI. Indeed, as will be shown in the following, some key experimental features of NBTI which are incompatible with the $P_h$ center and simple hole trapping models follow directly from the properties of the $E'$ center.
- The amount of positive charge visible and thus contributing to $\Delta V_{th}$ will depend on the position of the Fermi-level, that is, the gate voltage at which the degradation is monitored. For example, during OTF experiments, the Fermi-level is below the valence band edge and most defects will be positively charged (state 2 in Fig. 7) and thus visible. During recovery, the Fermi-level is moved towards mid-gap and a smaller fraction of the defects will be positive and visible. This occupancy effect will manifest itself as a change in the subthreshold slope, and is consistent with recent experimental results [16, 27, 44, 45].
- Since full annealing of oxide defects is only possible when the defect is neutral (state 3), defects having an electrically higher trapping level will show a slower recovery rate. Indeed, as will be shown, this is fully compatible with what is observed during...
NBTI recovery and explains the often observed bias dependence. Furthermore, this can explain the asymmetry between stress and recovery, with the recovery lasting considerably longer than the time required to build up the degradation.

- The fact that the precursor level is below the valence band and the defect level within the silicon bandgap is precisely what is expected of a defect responsible for NBTI. A higher energy level of the precursor, e.g. above the valence band, would cause most precursors to be already initially broken in PMOSFETs. The energy level of the created defect inside the silicon bandgap results in most defects to be positively charged (a donor-like defect) during both stress and recovery. A lower defect level would render the defects electrically neutral and thus not contributing to the threshold voltage shift.
- After NBTI stress an increase in $1/f$ noise has been reported [46, 47]. The prime suspect for $1/f$ noise are the $E'_0$ centers [29, 39], while $P_3$ centers do not create a suitable $1/f$ spectrum.

Although an extensive amount of literature is available on the $E'_0$ center and its qualitative behavior is well understood, no rate-equation based model that spans the full cycle of trap creation, recharging and discharging until final annealing seems to be available. As we will demonstrate in the following, however, for NBTI a full model is required in order to account for the asymmetry between stress and relaxation and the correct bias dependence of recovery.

VI. A Two Stage Model

Based on the above observations we formulate a new model for NBTI, where defect creation proceeds via a two stage process: In stage one, upon application of stress, holes can be trapped into near-interfacial oxygen vacancies via the MPE/MPFAT mechanism. In the second stage, the increased hole concentration considerably enhances interfacial oxygen vacancies via the MPE/MPFAT mechanism. In the stage one, upon application of stress, holes can be trapped into near-NBTI two stage process, where defect creation proceeds via a $1/f$ noise [49]. The total threshold voltage shift is thus given by

$$
\Delta V_{th}(t) = -\frac{\Delta Q_{ox}(t) + \Delta Q_{it}(t)}{C_{ox}}.
$$

In the following, detailed microscopic models for $\Delta Q_{ox}$ and $\Delta Q_{it}$ are developed.
Fig. 8: Qualitative degradation and recovery behavior predicted by our $\varepsilon'$ center model. Slightly artificial parameter values with a narrow distribution for $\Delta E_B$ where used to distinctly contrast the change in occupancy (transition from state 2 to state 3) versus true annealing (transition from state 3 to state 1). Left: The model can predict the asymmetric behavior during stress and recovery. In the above example, $B_1/B_2 \approx 2.5$ is obtained, just like in the real data of Fig. 4. Right: The model can predict the strong bias sensitivity during recovery. Dotted lines give the number of oxide defects (state 2 and 3), while the solid line corresponds to the positively charged center (state 2 only).

Based on the above rates, the rate equations describing the HDL switching trap are straightforward to set up. Regarding the rates we take the simplest possible case that can capture the currently available data. We assume that when the defect is in state 1 the trap energy lies at $E_T$, has a MPE barrier of $\Delta E_B$, and a MPFAT reference field $F_c$. When in state 2 and 3, the defect level is assumed to be at $E_T'$ with a small charging/discharging MPE barrier $\Delta E_C$. Although the barrier $\Delta E_C$ is expected to be considerably lower than the barrier $\Delta E_B$, it is responsible for the $\varepsilon'$ centers to act as 'slow states' in CP measurements. Nonetheless, for the data investigated here, $\Delta E_C$ can be neglected. We also neglect the MPFAT mechanism for charging/discharging since our data taken to sense these characteristics are recorded at relatively low fields. Transition from state 3 to state 1 (full annealing rather than electrical neutralization) proceeds over a barrier $\Delta E_A$.

The rate equations describing the transitions between the three states then read

\[ \frac{\partial f_1}{\partial t} = -f_1 k_{12} + f_3 k_{31}, \]
\[ \frac{\partial f_2}{\partial t} = +f_1 k_{12} - f_2 k_{23} + f_3 k_{32}, \]
\[ \frac{\partial f_3}{\partial t} = +f_2 k_{23} - f_3 k_{32} - f_3 k_{31}. \]

The probability of being in state $i$ is given by $f_i$ while the transition rates from state $i$ to $j$ are given by $k_{ij}$. Naturally, (9)–(11) are not linearly independent since the defect has to be in one of its three states ($f_1 + f_2 + f_3 = 1$). Being a linear equation system, solution of (9)–(11) is straightforward. The rates read as follows

\[ k_{12} = k_p^0(E_T, \Delta E_B, F_c) + k_p^0(E_T, \Delta E_B, F_c), \]
\[ k_{23} = k_p^0(E_T', \Delta E_C, 0) + k_n^0(E_T', \Delta E_C, 0), \]
\[ k_{32} = k_p^0(E_T', \Delta E_C, 0) + k_n^0(E_T', \Delta E_C, 0), \]
\[ k_{31} = \nu \exp(-\beta \Delta E_A), \]

where the shorthand $k$ (trap level, MPE barrier, MPFAT reference field) is used and $\nu \sim 10^{13}$ Hz is the typical attempt frequency for thermal transitions over energetic barriers. We remark that under conventional stress and recovery voltages the contribution of electrons is negligible. Only for positive bias which is applied in some of our recovery experiments they provide a significant contribution and impact the recovery dynamics, see Fig. 8.

In order to describe the response of a device to a change in the bias conditions, a certain number of defects $N$ is assumed to exist. Due to the amorphous nature of the Si/SiO$_2$ interface, each defect will be described by a unique configuration of random variables $X = (x, E_T, E_T', \Delta E_A, \Delta E_B, \Delta E_C)$. The joint probability density function is given by $g(X)$ and macroscopically observable quantities will be given by suitable averages

\[ \langle m \rangle = N \int dX \, mg(X). \]

For instance, the total positive charge contributing to $\Delta V_{th}$ is then given by the statistical average over the $\varepsilon'$ centers being in state 2, that is, positively charged, and we obtain

\[ Q_{ox}(t) = q\langle(1-x/t_{ox})f_2(t)\rangle, \]

with $t_{ox}$ the oxide thickness, and $\Delta Q_{ox}(t) = Q_{ox}(t) - Q_{ox}(0)$. The total number of oxide traps is given by all the defects not being in state 1 (the precursor state) and is simply

\[ N_{ox}(t) = 1 - \langle f_1(t) \rangle. \]

In order to keep the model as simple as possible, we will assume $g(X)$ to be given by independent homogeneous distributions in every variable, unless otherwise noted.

As the recovery in our ultra-thin and ultra-thick oxides looks basically the same, we conclude that the observed dispersion in time constants is primarily a property of the Si/SiO$_2$ interface. Consequently, the depth dispersion will be neglected ($x \approx 0$) to further simplify the model and thus $Q_{ox} = q(f_2)$. Naturally, a variation in $x$ will exist in reality, however, at the moment it does not appear necessary to include it into the model to explain our data.

Two interesting aspects of our $\varepsilon'$ center model are demonstrated in Fig. 8. First, it can predict asymmetric logarithmic stress and recovery behavior. Second, the model predicts a strong sensitivity to positive bias during recovery: Initially ($t_r < 10^{-15}$), after the stress bias is...
switched to the recovery voltage, most defects are positively charged (state 2) and nothing changes. Then, depending on the position of the Fermi-level during recovery, a possibly significant number of defects are electrically neutralized in the pico- and nano-second regime. In particular for positive bias, electrons from the conduction band cause a fast change in occupancy. It is important to realize that although the amount of visible positive charge can be small, the defect is not yet annealed and still there (not in state 1). With the artificially narrow distribution of $\Delta E_B$ used for demonstration purposes, real recovery only sets in at $10^{-3}$ with the recovery being fast when most defects are neutral (positive gate voltage) and slow when most defects are positive (negative voltage). With more realistic parameters, these transitions are blurred.

B. Coupled Interface State Generation

Once a hole has been trapped in the oxygen vacancy, a positive $E'$ center is obtained. One half of the $E'$ center is an unpassivated silicon dangling bond (DB). We now follow the arguments of Lenahan [35]: assume that in an unstressed device a certain number of hydrogen dangling bond (DB). We now follow the arguments of Lenahan [35]:

$\Delta E_B$ obtained in the current model by assuming that the H stays at $E'$ which is canonical over all to the interface, whose charge state quickly follows the Fermi-level in the substrate. The H has a non-zero probability of moving back to the $P_b$ center, thereby resetting the $E'/P_b$ complex to state 2, from which complete annealing is eventually possible. However, for typical stress conditions the full recovery would be outside the measurement window.

The rate equation describing such a $E'/P_b$ complex follow from a straight-forward extension of (9)–(11)

$$\frac{\partial f_1}{\partial t} = -f_1k_{12} + f_3k_{31},$$

(19)

$$\frac{\partial f_2}{\partial t} = +f_1k_{12} - f_2k_{23} + f_3k_{32} - f_2k_{24} + f_4k_{42},$$

(20)

$$\frac{\partial f_3}{\partial t} = +f_2k_{23} - f_3k_{32} - f_3k_{31},$$

(21)

$$\frac{\partial f_4}{\partial t} = +f_2k_{24} - f_4k_{42}.$$  

(22)

Again, the probabilities must fulfill $f_1 + f_2 + f_3 + f_4 = 1$ and one of the above equations must be omitted. The transition rates between state 2 and 4 are modeled in the spirit of [1,22] by thermal activation over a field-dependent barrier (cf. Fig. 9) as

$$k_{24} = ve^{-\beta(\Delta E_D - E_{3/2} - \gamma F)},$$

(23)

$$k_{42} = ve^{-\beta(\Delta E_D - E_{1} + \gamma F)},$$

(24)

Consequently, the probability of being in state 4 corresponds to a fixed positive charge at the $E'$ center and a depassivated interface state. However, only interface states above the Fermi level and up to mid-gap are assumed to be positively charged and their occupancy with an electron is given by $f_{it}(t)$, which is determined using the conventional SRH mechanism. Contrary to claims in [13], $f_{it}$ normally reaches its equilibrium value in the nanosecond regime (extending into the millisecond regime only for a switch of the Fermi-level below $E_G$ during stress to mid-gap) and thus $f_{it}$ may be set equal to the Fermi distribution in most cases.

Following [1,22,52], the barrier $\Delta E_D$ is assumed to be given by a narrow Gaussian distribution ($\sigma \approx 100\mathrm{meV}$), while the energy-levels $E_i$ of the electrically active states are assumed to be homogeneously distributed over the lower-half of the silicon bandgap. We remark that a strong correlation between these two distributions is to be
expected, that is, $E_q$ should also be given by two Gaussians in the bandgap [43]. Also note that in contrast to the $E'$ center model, repassivation of interface states is assumed to be independent of their charge state, which might not be correct. We thus consider the above model a first-order approximation with the minimal number of free parameters. More detailed experimental data would be required to justify a refined version of the model.

The random variables of our $E'/p_{pH}$ complex are thus $X = (x, E_T, E'_T, \Delta E_A, \Delta E_B, \Delta E_C, \Delta E_F)$ and their joint probability density function $g(X)$ used in the statistical average (16) is constructed from independent distributions.

The amount of positive charge stored in the $E'$ part of the complexes is now given by the statistical average of complexes being either in state 2 or 4, and we have

$$Q_{ox} = q\langle(1-x/t_{ox})(f_2 + f_4)\rangle$$ (25)

The amount of positive charge stored in the interface states is given by the average of the probability of having depassivated DBs ($f_4$) times the probability that the created electrical level is unoccupied (1 $- f_0$) and for $E_q$ within the lower-half of the silicon band-gap (the donor-like states) given by

$$Q_{it} = q\langle f_4(1 - f_0)\rangle,$$ (26)

while the total number of available interface states is simply obtained from $N_{it} = \langle f_4 \rangle$. For the data analyzed here, the occupancy of the created interface states is of minor relevance since normally both during stress and recovery the Fermi-level is close to the silicon valence band ($f_0 \approx 0$) and consequently most interface states will be positively charged and thus $Q_{it} \approx qN_{it}$.

In order to explain the presently available data, it is sufficient to use a reduced set of random variables and we use $x \approx 0, \Delta E_A \approx \Delta E_B,$ and $\Delta E_C \approx 0$ in the following examples. Further simplification, like $E'_T \approx E_T$, significantly impacts the quality of the model as will be discussed later.

A typical evolution of the densities of the precursors, the created $E'$ centers, and the created interface states is shown in Fig. 10 during stress and after 10s of recovery. Although uniform distributions are used in the model, the resulting DOS of $E'$ centers is non-uniform. This is because states closer or below the Fermi-level are electrically neutral and thus have a larger annealing rate. Consequently, the higher the trap level $E'_T$, the longer it will take the defect to fully anneal, consistent with the interpretation given in [27].

VII. COMPARISON WITH MEASUREMENTS

Simulation results for the SiON devices are compared to the data of Fig. 3 and Fig. 4. The simulated prefactors during stress are in very good agreement with the data, see Fig. 11. Details of the simulation demonstrate also that the model can reproduce the asymmetry between stress and recovery, see Fig. 12. We remark that this is the first time that a model can reproduce both OTF and recovery data in a wide temperature and voltage range, providing a theoretical link between these two measurement techniques.

A particularly challenging data set is given in Fig. 13, where five devices are brought to the same level of degradation. During recovery, various bias switches are used to probe both the occupancy effect (amount of charge visible depending on the Fermi-level) as well as the impact of the occupancy on the recovery dynamics. Again, very good agreement between theory and data is obtained.

Results of similar accuracy have been obtained for the HK device, see Fig. 14. Compared to the SiON-k devices, a considerably larger amount of $E'$ centers is created.

This is in contrast to data taken on thick SiO$_2$ devices where a considerably smaller amount of $E'$ centers is created, see Fig. 15. Finally, we test the underlying physical assumptions of the model by applying a recently developed poly-heater technology allowing rapid switches of the device temperature [20]. The appealing feature of this technique is that devices can be brought to the same stress level from which recovery can be monitored at different $T$. As shown in Fig. 16, these insightful experimental data are well reproduced by the model, confirming its validity.

VIII. ALTERNATIVE MODELS AND WHY THEY DO NOT WORK

One question that naturally needs to be raised is whether really all features of the $E'$ center are needed in order to explain NBTI. In a nutshell, the answer is simply ‘yes’ and we will demonstrate in this section why this is the case. We have to keep in mind, however,
that the following allegedly simpler models are incompatible with the known properties of the $E'$ center. Rather, we wish to show that each aspect of the model has its correspondence in experimental data.

### A. Occupancy Effect

First, we demonstrate the fundamental impact of the trap occupancy on the simulated stress and recovery characteristics. From an electrical point of view, the occupancy effect (transitions between state 2 and 3) is responsible for a change in the subthreshold slope. It also explains the asymmetry between the degradation and recovery dynamics. A simplified model that neglects state 3 (the electrically

### B. Level Shift

The next example effectively demonstrates what happens when the level shift is neglected, that is, when $E'_L = E_F$ is assumed. In order to have a stable precursor configuration, $E_F$ must lie well below the silicon valence band. During stress, holes are trapped at that level. During relaxation the holes are kept in that low level only by the MPE barrier and must go back to the valence band, a process independent of the intermediate bias switch is observed. This procedure was repeated for increasing durations of bias switches (10, 40, 150, 500s). Simulation results are given by the lines, which show very good agreement with the data, capturing both the occupancy effect (evenly numbered panels) and the acceleration/retardation of recovery as a response to the gate bias.
occur via a prior recovery the temperature is quickly switched to −40°C. Finally, in a fourth device the temperature is first switched to −40°C, then after 1s recovery to +40°C, and finally after 100s to 125°C. Excellent agreement of the model with data is again obtained.

IX. CONCLUSIONS

We suggest a two stage model for the negative bias temperature instability based on established properties of E′ centers. Creation of E′ centers from their oxygen vacancy precursors is suggested to occur via a multiphonon-field-assisted hole trapping mechanism. The created E′ centers are then suggested to favor the depassivation of interface states, which results in a coupling of created oxide and interface state component.

This model can explain degradation and recovery over a wide range of bias voltages and stress temperatures, the observed asymmetry between stress and recovery, and the strong sensitivity to bias and temperature during recovery. Excellent agreement with data from three vastly different technologies (thick SiO2, SiON, and HK) is obtained, supporting the idea that NBTI is determined by the chemistry of the amorphous SiO2/Si interface region. The model has the minimum number of parameters required to explain the experimentally observed features of NBTI.

X. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community’s Seventh Framework Programme under grant agreement n°216436 (project ATHENIS). We also gratefully acknowledge stimulating discussions with P. Lenahan and A. Shluger.

REFERENCES


Fig. 18: Characteristic behavior of a model that neglects the occupancy effect and does not consider the level shift (no state 3 and $E_0 = E_T$). Stress and recovery are always symmetric (not shown). Devices stressed at the same voltage show practically no voltage dependence during recovery.


