The Time Dependent Defect Spectroscopy (TDDS) for the Characterization of the Bias Temperature Instability

T. Grasser*, H. Reisinger•, P.-J. Wagner*, F. Schanovsky*, W. Goes*, B. Kaczer°

*Institute for Microelectronics, TU Wien, Austria • Infineon, Munich, Germany ° IMEC, Leuven, Belgium

***Abstract*— We introduce a new method to analyze the statistical properties of the defects responsible for the ubiquitous recovery behavior following negative bias temperature stress, which we term time dependent defect spectroscopy (TDDS). The TDDS relies on small-area metal-oxide-semiconductor field effect transistors (MOSFETs) where recovery proceeds in discrete steps. Contrary to techniques for the analysis of random telegraph noise (RTN), which only allow to monitor the defect behavior in a rather narrow window, the TDDS can be used to study the capture and emission times of the defects over an extremely wide range. We demonstrate that the recoverable component of NBTI is due to thermally activated hole capture and emission in individual defects with a very wide distribution of time constants, consistent with nonradiative multiphonon theory previously applied to the analysis of RTN. The defects responsible for this process show a number of peculiar features similar to anomalous RTN previously observed in nMOS transistors. A quantitative model is suggested which can explain the bias as well as the temperature dependence of the characteristic time constants. Furthermore, it is shown how the new model naturally explains the various abnormalities observed.**

I. Introduction

The lifetime of metal-oxide-semiconductor field effect transistors (MOSFETs) is limited by unavoidable degradation during operation. One of the most critical degradation mechanisms in p-channel MOSFETs is the negative bias temperature instability (NBTI). As the name implies, it is observed when a pMOSFET is subjected to negative bias at the gate with the other terminals grounded. The degradation is considerably accelerated at elevated temperatures. Although the detrimental impact of negative bias temperature stress (NBTS) has been known for more than 40 years and is of highest relevance to the semiconductor industry, the detailed physics behind the phenomenon are still highly controversial [1]. The most popularized explanation of this phenomenon invokes a reaction-diffusion (RD) driven depassivation of hydrogen-passivated interface states [2, 3], initially obscured by elastic (temperature-independent) hole trapping in the oxide [4].

The first clues to the limitations of the RD theory were found when the recovery of the degradation was studied in greater detail [5–8]: most importantly, it was found that the degradation consists of a recoverable component on top of a nearly permanent part [7]. Also, recovery starts on time scales faster than accessible by experiment (below microseconds), continues longer than is usually measured (up to weeks), and is strongly sensitive to the applied gate bias. This is inconsistent with a diffusive process involving neutral hydrogen species but is intuitively compatible with a hole trapping mechanism with widely distributed time constants [9, 10].

In a first attempt to explain experimental data, simple elastic hole trapping models were used [7] which explain the wide distribution of time constants by assuming that hole traps are spatially distributed into the depth of the oxide. Due to the necessary tunneling transition from the channel, traps deeper in the oxide have larger time constants [11]. However, for the ultra-thin oxides employed in modern devices having thicknesses of 1-2 nm, the maximum time constants explainable with such a model are in the order of milliseconds [11]. Yet, experimentally observed recovery continues unabated for weeks and months, seemingly linear on a semi-log scale [5, 8, 12].

Another phenomenon where the assumption of a wide distribution of time constants is commonly invoked is the $1/f$ noise [13–15]. In small-area devices only a handful of defects are present and their time constants can be experimentally assessed from the dispersion of random telegraph noise (RTN) properties [15]. These time constants are temperature activated and can be very sensitive to the applied gate bias. In particular, models originally developed for RTN and $1/f$ noise qualitatively capture many features typically associated with NBTI [10]. As such it has been speculated that the defects responsible for RTN are also the cause of NBTI [9, 10]. Like $1/f$ noise, NBTI is typically studied on large-area devices, where a large ensemble of defects acts simultaneously. Consequently, the individual defect properties are averaged out and unambiguous identification of the physics is difficult. In the following, we use small-area devices, which allow us to clearly and unambiguously identify and study the individual defects constituting the macroscopically observable behavior. In order to guarantee that the discrete steps observed in the recovery traces of small-area devices [9, 16] are truly the main ingredient of the log-like recovery ubiquitous in large area devices [9, 17], we have compared the average of 25 small-area device recovery traces to a single large-area device of comparable effective width, with favorable outcome [18].

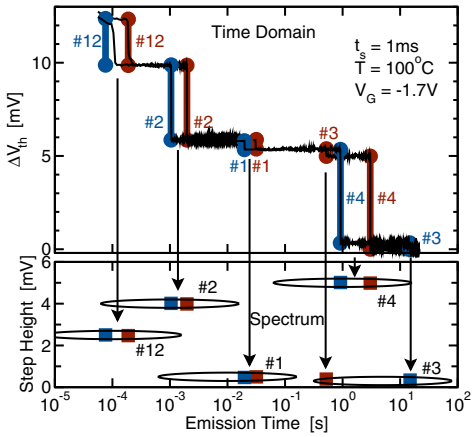

Fig. 1: Two typical threshold voltage recovery traces of a previously stressed small-area pMOSFET. The measured data are given by the (slightly noisy) thin black lines in the top part of the figure. The thick blue and red lines together with the symbols mark the automatically extracted emission times and step heights which are unambiguous fingerprints of each defect and build the spectral map (bottom).

II. The Time Dependent Defect Spectroscopy

We proceed by introducing an intuitive analysis method, which we term *time dependent defect spectroscopy* (TDDS). It uses small-area devices, where recovery after stress proceeds in discrete steps occurring at stochastic times [9, 16, 18–20]. For details of the measurement procedure see [18]. We conclude that just like in RTN, each discrete step is due to the emission of a single trapped carrier [16]. As such, the TDDS is similar in spirit to the successful deep-level transient spectroscopy (DLTS) technique [21], which has also been applied to small devices [22]. However, rather than assuming that after application of a charging pulse for a certain amount of time

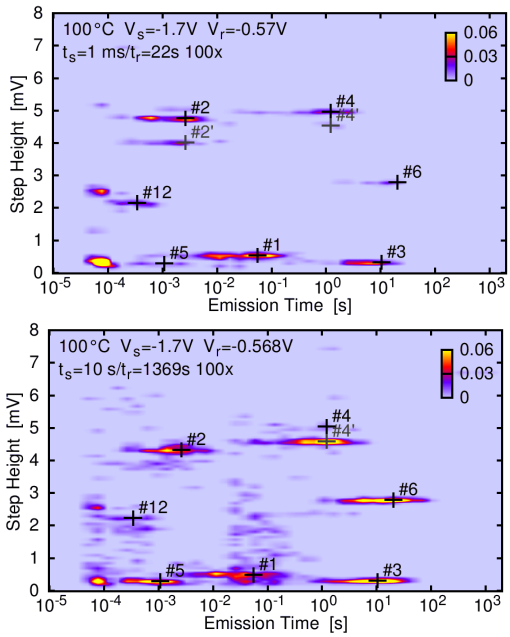

Fig. 2: Spectral maps at two stress times, $t_s = 1\,\text{ms}$ (top) and $t_s = 10\,\text{s}$ (bottom). With increasing stress time, the number of defects in the map increases as defects with $\bar{\tau}_c \lesssim t_s$ have a significant probability of being charged after stress. Note the increased noise level (hits outside clusters) although the same number of defects constitute the map. The width of each cluster is given by the exponential distribution of τ_e and the extracted $(d, \bar{\tau}_e)$ are marked by '+'.

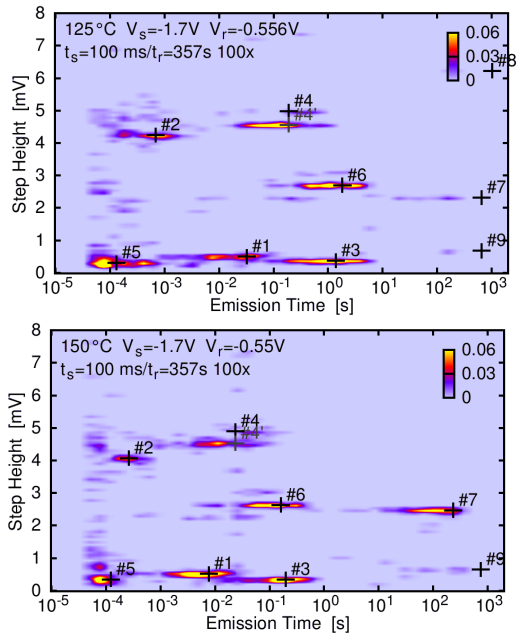

Fig. 3: Temperature dependence of the spectral maps after $t_s = 100\,\text{ms}$. With increasing temperature, the emission times decrease. As also the capture time constant decreases, the clusters associated with each defect appear after a shorter stress time at higher temperatures. Defects #2 and #5 partially move outside our experimental window, while #7, #8, and #9 enter. Nonetheless, the parameters of #7, #8, and #9 cannot be reliably extracted at these temperatures.

all defects are fully charged [22], we exploit the fact that the capture time constants are widely distributed, an issue also observed in DLTS literature as a non-saturating behavior in the DLTS spectra [23]. It is particularly this time-dependence of the TDDS spectral maps that allows for a detailed assessment of the physical phenomenon.

For the TDDS, a number of stress/recovery experiments has to be performed for a reliable characterization of the stochastic process. For this, the device is repeatedly stressed for the same amount of time and the subsequent recovery trace is recorded. The statistical properties of the discrete steps in the recovery traces are then analyzed by collecting the heights d and emission times τ_e of each emission event, see Fig. 1. The collected pairs (τ_e, d) are then binned into a 2D histogram, see Fig. 1 (bottom).

As has been demonstrated in RTN literature [15, 24], the step height observed in the threshold voltage shifts does not depend on the microscopic physics of the defect but rather on the interaction between the lateral position of the defect and the random potential inside the channel [24] and thus allows for the identification of individual defects. Step heights larger than the value obtained from the charge-sheet approximation, $qt_{ox}/(\varepsilon WL)$ with t_{ox} being the oxide thickness, ε the oxide permittivity, W and L the device width and length, indicate a defect in a so-called current percolation path [25] and is commonly referred to as a 'giant' step in RTN literature [15, 24]. For the device used in this study, the charge sheet approximation gives about 0.9 mV, while experimentally step heights of more than 6 mV have been observed in our samples. For even higher values and details on the distribution of the step heights see [26].

In order to gather sufficiently accurate statistics, the stress and recovery cycle is repeated $N = 100$ times and the entries in the 2D histogram are normalized by N to obtain the spectral map after the stress time t_s. The whole procedure is then repeated M times with increasing stress time, where we used $t_s \in \{1\,\mu\text{s}, 10\,\mu\text{s}, \ldots, 10\,\text{s}\}$. The obtained set of M spectral maps associated with a particular stress and recovery condition is then used to thoroughly analyze the defects.

A. Example Spectral Maps

Two examples for such spectral maps obtained from a production quality pMOSFET with a 2.2 nm thick plasma-nitrided oxide [8, 18] stressed at $-1.7\,\text{V}$ and $100\,^\circ\text{C}$ are shown in Fig. 2. Clearly, marked clusters of (τ_e, d) pairs evolve, the intensity of which typically follows $P_c = 1 - \exp(-t_s/\bar{\tau}_c(V_{\text{stress}}))$, with $\bar{\tau}_c(V_{\text{stress}})$ as the average capture time at the stress bias $V_G = V_{\text{stress}}$. This expression for P_c is valid as long as $\bar{\tau}_c(V_{\text{stress}}) \ll \bar{\tau}_e(V_{\text{stress}})$, which is mostly the case. Should $\bar{\tau}_c(V_{\text{stress}}) \sim \bar{\tau}_e(V_{\text{stress}})$ the defect produces RTN making the extraction more problematic as $P_c(t_s \to \infty) = \bar{\tau}_e/(\bar{\tau}_e + \bar{\tau}_c) < 1$, requiring the extraction of both $\bar{\tau}_c(V_{\text{stress}})$ and $\bar{\tau}_e(V_{\text{stress}})$. The position on the emission time axis remains fixed with stress time while the step height is subject to slight variations, depending on the most dominant current conduction path [24, 25] which can change with the occurrence of additional charged defects. Each cluster corresponds to an individual defect and is labeled accordingly in the maps.

When the procedure is repeated at higher temperatures, the same clusters are observed but clearly have shorter emission times, see Fig. 3, confirming that we are dealing with a thermally activated emission process. Also, all clusters appear already after shorter stress times, demonstrating the thermally activated nature of charge capture. For example, after a 100 ms stress the cluster belonging to defect #7 is already fully developed at $150\,^\circ\text{C}$ while it only gradually begins to form at $125\,^\circ\text{C}$. Recall that the spectral maps show the emission times at the recovery voltage V_{relax}, while P_c is controlled by $\bar{\tau}_c$ and $\bar{\tau}_e$ at V_{stress}. This is in contrast to RTN measurements where the capture and emission times are always recorded at the same bias condition.

B. Theoretical Considerations

For the analysis of the spectral maps we need to understand what theory predicts. In order to capture the stochastic nature of the trapping and detrapping process, the model equations, which are given by coupled partial differential equation systems, are solved

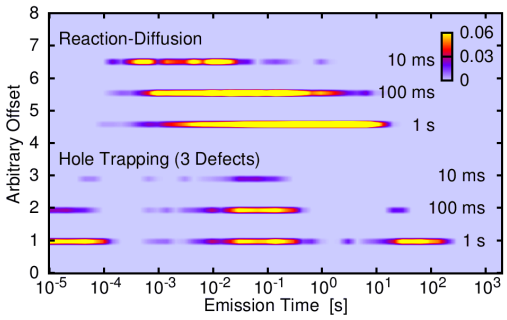

Fig. 4: Theoretical spectral maps obtained from the stochastic solution of the reaction-diffusion model and a simple hole-trapping model for three stress times. The top part of the figure shows the solution of the RD model which produces a single moving cluster, which has a maximum at $\tau_e \sim t_s$ and increases in intensity, in stark contradiction to the experimental data. The lower part shows a solution of a hole-trapping model, which perfectly resembles the experimentally observed behavior.

using the Stochastic Simulation Algorithm [20, 27]. As can be seen in Fig. 4 (top), the conventional RD model predicts the occurrence of relatively wide clusters in the spectral map where 80% of the emission events occur in about 3.8 decades around the average emission time [20] compared to the 1.3 decades observed experimentally. Furthermore, the cluster resulting from a diffusion-limited process would move to larger emission times for increasing stress time, while our experiments only reveal clusters fixed with stress time. And finally, this RD cluster should increase in magnitude as a function of $t_s^{1/6}$ [1], while the maximum amplitude of the charge trapping clusters is obtained from the case $P_c = 1$, corresponding to a 100% capture probability. As a consequence, the RD cluster would begin to dominate the spectral maps at a very early time. In addition, the initial charge trapping step would lead to temperature independent clusters in the spectral maps while not a single temperature-independent cluster has been observed. In short, not a single prediction of the stochastic RD model is consistent with our experiments.

In contrast, a simple hole capture and emission model produces narrow clusters with fixed emissions times for a particular recovery condition (cf. Fig. 4, bottom), in excellent agreement with the experimental data. While this feature can be reproduced by any first-order reaction model, the bias and temperature dependencies of the capture and emission time constants is crucial. As a reference we recall the frequently used model based on a simple nonradiative multiphonon (NMP) process, as used for instance by Kirton and Uren [15]

$$\bar{\tau}_c = \tau_0 \, e^{\beta \Delta E_B} \frac{N_V}{p}, \tag{1}$$

$$\bar{\tau}_e = \bar{\tau}_c \, e^{\beta E_{TF}} \approx \tau_0 \, e^{\beta(\Delta E_B + \Delta E_T)} \, e^{xF/V_T}, \tag{2}$$

with $\tau_0^{-1} = N_V \, v_{th} \, \sigma_0 \, e^{-x/x_0}$ and $E_{TF} = E_T - E_F$. Assuming to first-order that the oxide is charge free, the oxide trap level depends on the applied bias via $E_T = E_{T_0} - q\varphi_s + qxF$, which gives $E_{TF} = E_{VF} + \Delta E_T + qxF$, with $\Delta E_T = E_{T_0} - E_{V_0}$. F is the modulus of the oxide field, p the surface hole concentration, ΔE_B the (bias-independent) NMP barrier, $\beta^{-1} = k_B T$, $V_T = k_B T/q$, x the distance of the defect into the oxide, v_{th} the thermal hole velocity, φ_s the surface potential, x_0 from a simple WKB approximation for a large and thin barrier of height ϕ as $x_0 = \hbar/(2\sqrt{2m\phi})$, and σ_0 the capture cross section. Note that in this model $\bar{\tau}_c$ only weakly depends on bias as $1/p$, while depending on the trap depth x, a rather strong exponential dependence on F is predicted for $\bar{\tau}_e$. However, around V_{th}, F depends only weakly on V_G, and so does $\bar{\tau}_e$.

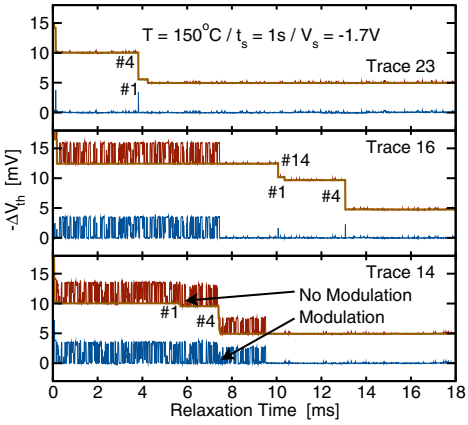

Fig. 5: In addition to the discrete recovery steps, the traces can contain temporary RTN, which disappears after about the average slow emission time, $\bar{\tau}_e^s$, of the defect, see the three selected traces above. The extracted step heights (brown) are subtracted from the raw data (red) to yield the noise trace (blue). Just like with any other defect, the tRTN step height can be modulated by a change in charge in another defect, in the above example defect #4. By contrast, the emission event associated with #1 does not modulate the step height. Although the tRTN defect has a step-height and average (slow) emission time similar to #4, it is a different defect as emission events of #4 are completely uncorrelated (middle and bottom). Also, the step height of the tRTN defect has a different bias dependence.

Since such a charge trapping mechanism is a Poisson process, just like charge capture and emission leading to RTN [15], the capture and emission times are exponentially distributed. Consequently, the probability of τ_e falling into bin i is given by $P_e = P(\tau_i \le \tau_e < \tau_{i+1}) = \exp(-\tau_i/\bar{\tau}_e) - \exp(-\tau_{i+1}/\bar{\tau}_e)$. However, in order for the defect to be discharged, it has to be previously charged after the stress time t_s, the probability of which is given by P_c. In total, the probability of having an event in bin i is given by $P = P_c P_e$. By fitting the set of spectral maps to $P(\tau_c, \tau_e)$, the characteristic parameters $\bar{\tau}_e$ and $\bar{\tau}_c$ as well as the step height d can be extracted for each cluster.

C. Peculiarities

Although a first-order reaction can describe many clusters with very good accuracy, deviations from this behavior occur quite regularly. The most commonly observed feature is the splitting of the cluster into two peaks due to the electrostatic interaction when another trapped charge in the same percolation path modulates the step height of the defect. More complex deviations, like different emission times for the two peaks, have also been observed but will be discussed in a separate publication.

While the spectral maps are dominated by defects which discharge in the expected manner, RTN appeared on a few traces in our particular device after a stress time of about 1 s. Quite intriguingly, the RTN disappeared after a certain time, for instance after about a few milliseconds at 150 °C. We consequently suggest to term this phenomenon *temporary RTN*, or tRTN. The occurrence of this RTN was found to follow a stochastic process, similar to normal charge capture, that is, with increasing stress time and increasing T the number of traces showing tRTN increased. Also, the tRTN capture and emission times showed the same temperature and bias dependence as known from normal RTN.

Consequently, four characteristic time constants are required for such a defect: the *slow capture time constant* $\bar{\tau}_c^s$ initiating the tRTN process, the *fast capture and emission time constants* $\bar{\tau}_c^f$ and $\bar{\tau}_e^f$ determining the on and off times of the tRTN process, and the *slow emission time constant* $\bar{\tau}_e^s$ eventually terminating the tRTN signal.

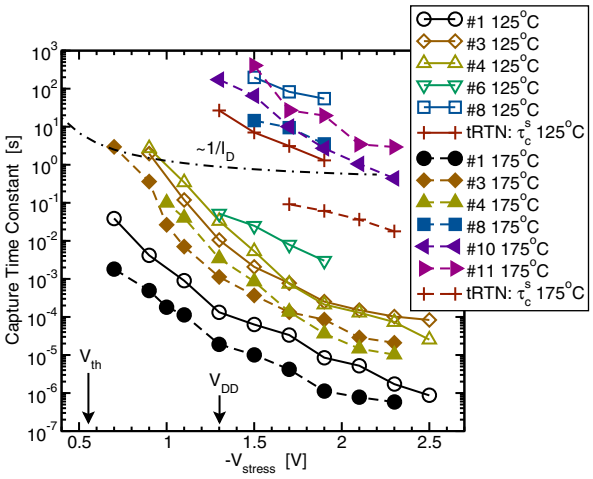

Fig. 6: Voltage and temperature dependence of the capture time constant for 7 defects. Defect #6 was visible during the initial experiments only (taken at 125 °C) and then disappeared permanently, #10 and #11 were outside our experimental window at 125 °C. A strong field/voltage dependence slightly different from exponential is observed for all defects. Also shown is the $1/I_D$ tendency expected from the standard SRH-like models which is considerably weaker.

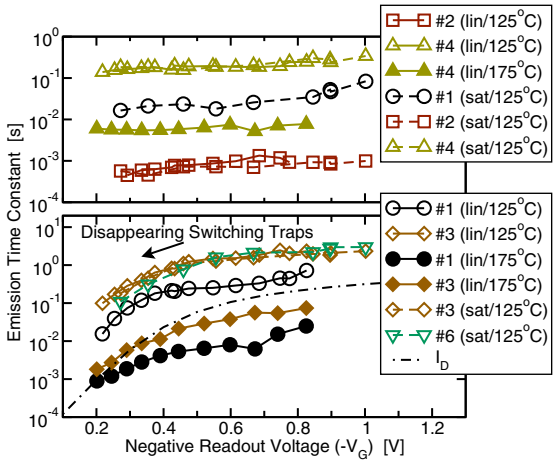

Fig. 7: Gate voltage dependence of the emission time constants at 125 °C and 175 °C, measured in the linear and the saturation regime. Defects #1/sat, #2, and #4 (top) may be also described by a standard model such as (2) because they show only a weak (if at all) field dependence at lower voltages. In contrast to that, $\bar{\tau}_e$ of defects #1/lin, #3, and #6 (bottom) shows a pronounced V_G dependence at low bias. This behavior seems to coincide with the interfacial hole concentration, which is roughly proportional to I_D. Such a behavior cannot be explained with a standard model.

An example for the occurrence of tRTN is shown in Fig. 5. Following the construction of the discrete approximation of ΔV_{th}, the discrete signal is subtracted from the original trace to yield the *noise trace*. As can be seen in Fig. 5, tRTN is visible from the onset of recovery and disappears on average after $\bar{\tau}_e^s \sim 8$ ms. Just like the step height of 'normal' defects, the step height or tRTN can be modulated by the charge state of other defects, in this case by the emission of charge from #4 but not by #1. The time constants of the tRTN defect show a very strong bias dependence, that is, tRTN is only visible in a narrow window. Also, tRTN shows a marked temperature dependence in all four characteristic time constants.

Also interesting is the observation that particularly heavier stresses resulted in defects *disappearing* (observed regularly for #6, #7, and #8) for a certain period. For example, #7 was inactive for more than a month during measurements at 125 °C after having been stressed a 100 times at −1.7 V/175 °C for 10 s. Contrary to that, defect #1 was present during the whole duration of the study and only disappeared once for two hours after having been stressed at −1.7 V/75 °C for 10 ms. A similar behavior was previously observed in RTN and was explained by the defect having an additional metastable state [28].

We particularly note that defects seem to disappear from the map without any impact on the threshold voltage shift, meaning there is no permanent degradation directly linked to their disappearance.

III. Extraction of Defect Parameters

To study the bias and temperature dependence of the basic defect parameters we recorded 80+ sets of spectral maps (with 5 to 8 maps per set) by continuously measuring on a single device over a period of three months and for a few weeks after a three months break. Quite intriguingly, the spectral maps remained basically the same for the whole duration. Although the individual traces show an increasing offset with increasingly heavier stress, probably due to a permanent degradation component, the spectral maps remained basically unaffected by this offset.

Within our range of voltages, temperatures, stress and relaxation times, 13 defects could be identified on this device and the bias and voltage dependence of their parameters extracted, at least to a certain degree. Defects that stayed within our experimental window and did not disappear could be studied over the whole operational regime of the transistors, starting from slightly above threshold up to close to oxide breakdown.

As an example, the extracted capture time constants are shown in Fig. 6, which are clearly temperature activated with an activation energy of about 0.6 eV and depend only approximately in an exponential manner on the stress bias. As has been reported for RTN [15], this is incompatible with the $1/p \sim 1/I_D$ dependence of the SRH-like model (2). In fact, for all defects which could be traced over a wider window, $\log(\bar{\tau}_c)$ also showed a marked non-linear field dependence which can be fit by a quadratic relation $-c_1F + c_2F^2$. It is particularly worthwhile to point out that this non-linearity is different for all defects and must as such be related to the details of the defect properties as in this regime both F and p depend linearly on V_G. Also note that were these defects studied only within the typical RTN measurement window, this non-linearity would have gone unnoticed.

From a reliability point of view this non-linearity in $\log(\bar{\tau}_c)$ is of fundamental importance as it connects the degradation accumulated in an accelerated stress regime to the degradation level expected under operating conditions. The observed non-linearity in the individual defects is also consistent with the observed non-linearity in macroscopic experiments, where a $\Delta V_{th} \sim \log(\bar{\tau}_c)$ behavior could be expected from averaging a large number of defects [10]. Accurate back-extrapolation to operating conditions is thus only possible when this non-linearity is properly accounted for.

Also of particular practical interest is the bias dependence of $\bar{\tau}_e$, as a strong bias dependence would allow one to efficiently remove the charges from the oxide by applying a small positive bias. In fact, a strong bias dependence of the macroscopically measured ΔV_{th} has been previously observed [7, 12, 29], albeit of a more complicated form than previously anticipated. It is worthwhile to recall that this issue was one of the first problems noted regarding reaction-diffusion theory [12], as the assumed diffusion of a neutral species is difficult to reconcile with any bias dependence.

Fig. 7 shows $\bar{\tau}_e(V_G)$ for five selected defects as measured in the linear and saturation regime. Defects #2 and #4 are insensitive to changes in V_G and have the same $\bar{\tau}_e$ in both regimes. On the

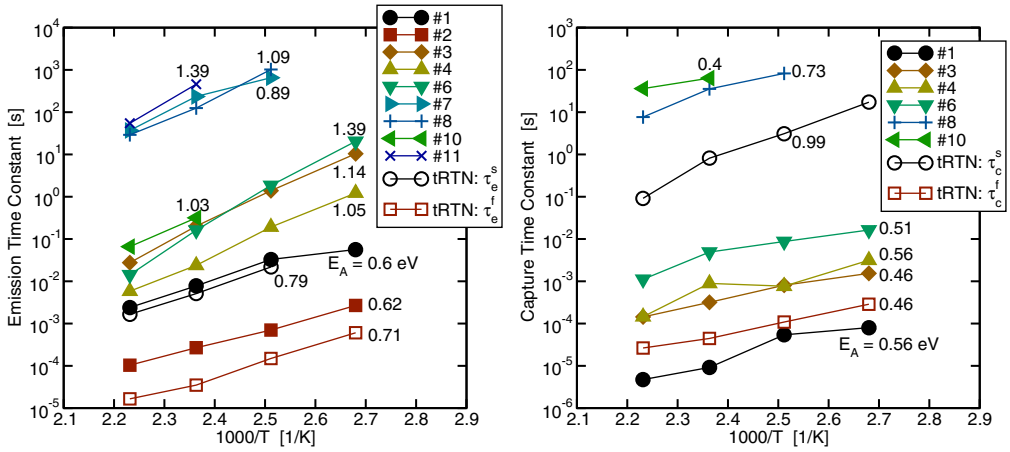

Fig. 8: Left: Arrhenius plot of the emission time constant $\bar{\tau}_e$ for a recovery bias of $-0.55\,\mathrm{V}$. The approximate activation energies are given for each defect where a wide spread is observed. **Right**: Arrhenius plot of the capture time constant $\bar{\tau}_c$ for a stress bias of $-1.7\,\mathrm{V}$. While the extraction of $\bar{\tau}_c$ is less reliable compared to the extraction of $\bar{\tau}_e$, the spread in activation energies seems to be smaller.

other hand, defects #3 and #6 demonstrate a very strong sensitivity when $|V_G|$ moves below $|V_{th}|$. Interestingly, this regime coincides with a large drop in the interfacial hole concentration. As with the bias dependence of $\bar{\tau}_c$, this is incompatible with a SRH-like model which predicts only a weak bias dependence in this regime. We have previously associated such a behavior with a switching trap [10] which differs from a 'normal' trap by the fact that, once created, its charge state can be controlled by the Fermi-level in the substrate prior to annealing. The most interesting case is thus defect #1, which in the saturation regime only shows a weak bias dependence but demonstrates switching trap behavior in the linear regime. The simplest conclusion to draw from this is that in fact both types of behavior are special cases of one and the same general defect type.

The temperature dependencies of $\bar{\tau}_e$ and $\bar{\tau}_c$ are shown in Fig. 8, demonstrating a wide spread in the activation energies, with the spread in $\bar{\tau}_e$ appearing larger than that observed in $\bar{\tau}_c$. As also observed in Fig. 6, the defect responsible for tRTN appears in no way different compared to the other defects, leading us to conclude that tRTN is just another special form of the general defect type.

A. The TDDS Measurement Window

Typical RTN measurements require the capture and emission times to be about equal and within the measurement window. Furthermore, the measurement becomes difficult to deconvolute when there are contributions from more than a single defect. The situation is fundamentally different and considerably improved for the TDDS:

- The basic requirement is only that the time constants fall within the experimental window. Typical time windows would be from $10\,\mu\mathrm{s}$ up to 100 or even 1 ks (covering about 8 decades in time), where the upper bounds are mostly "limited by the experimenters' patience" [30]. Occasionally, a defect may have $\bar{\tau}_c$ outside the experimental window while $\bar{\tau}_e$ can be fully characterized. In our particular device this is the case for defects #2, #5, and #12, depending on the temperature and bias: Due to the temperature dependence of the time constants, adjustment of the temperature allows to shift defects into the window. Similar considerations hold for $\bar{\tau}_c$, which can be brought into the window by appropriately selecting the stress voltage and temperature.
- Even when more than a dozen defects contribute to the spectral map, each of them can be analyzed provided that their capture times are separated by about a decade and the step height by more than about 0.5 mV (for the current device). The latter can often be enforced by suitably adjusting the drain bias during recovery, which may significantly impact the step height. For example, defects #1 and #3 have a step height of about 0.5 mV at $V_D = -1.2\,\mathrm{V}$. At $|V_G| < |V_{th}|$, the much stronger bias sensitivity of #3 results in a near overlap of the respective clusters on the spectral map. In contrast, at $V_D = -0.1\,\mathrm{V}$ the step heights are 0.3 mV and 1.5 mV and no overlap occurs.
- The gate bias dependence of the capture time constant can be measured over nearly the whole operation range of the transistor, from slightly above threshold (in order to provide a sufficient probability of charging the fastest defects) up to oxide breakdown.
- The gate bias dependence of the emission time constant is somewhat restricted since at low $|V_G|$ the minimum time constant obtainable with our equipment can increase significantly due to the small drain current. On the other hand, at high $|V_G|$, about 1 V in the present example technology, the drain current becomes too large to allow for a sufficiently accurate extraction of ΔV_{th}.

The strength of the TDDS becomes apparent from Fig. 9, where the capture and emission times of defect #1 are shown as a function of $|V_G|$. Using the conventional RTN analysis method, this defect could have only been monitored in the rather narrow window where $\bar{\tau}_c \sim \bar{\tau}_e$, not to mention the tremendous difficulties caused by the interference of other defects. Interestingly, when measured in the saturation regime $\bar{\tau}_e(V_G)$ is practically bias independent for $|V_G| < |V_{th}|$, while in the linear regime the switching trap behavior becomes apparent.

IV. Model

As has been noted previously, a number of features in $\bar{\tau}_c$ and $\bar{\tau}_e$ have been revealed by the TDDS which are inconsistent with standard theories:

- A non-linear bias dependence of $\log(\bar{\tau}_c)$, which is different for every defect. Here we remark that standard NMP theories predict a linear field dependence [31] for the strong electron-phonon coupling case, so the reason for this discrepancy must be investigated.
- A strong bias dependence of $\bar{\tau}_e$ for $|V_G| < |V_{th}|$ in some defects, depending on the drain (and probably substrate) bias. Such a strong bias dependence is likely due to the existence of a metastable state [20].

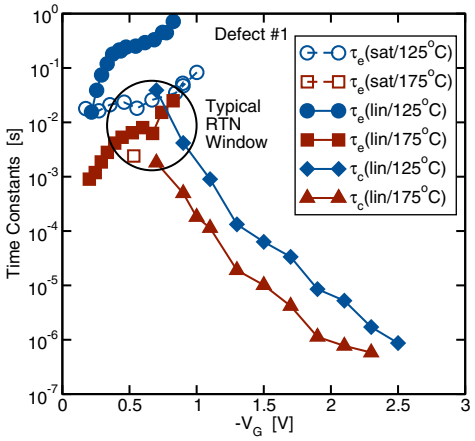

Fig. 9: The TDDS allows the measurement of the capture and emission times over an extremely wide range. Shown are also the emission times measured in the linear ($V_D = -0.1\,V$) and saturation regimes ($V_D = -1.2\,V$). For comparison a typical RTN measurement window is shown which requires $\bar{\tau}_e \sim \bar{\tau}_c$. When measured in the linear regime, the defect acts like a switching trap, while in the saturation regime this feature is not visible.

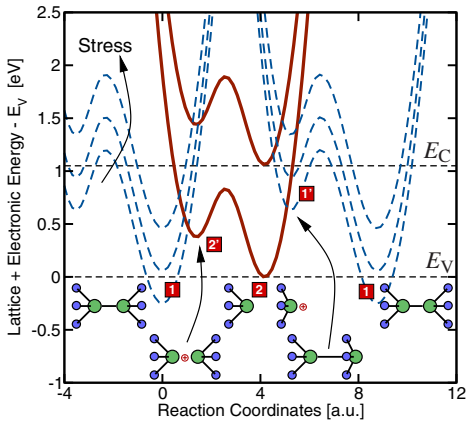

Fig. 10: Configuration coordinate diagram showing the adiabatic total energy potentials corresponding to the four defect configurations. The energy is given relative to the valence band edge E_V. In state 1 and 1' the electron is at the defect site, that is, the electronic contribution to the total energy shifts the potential up with increasing NBTS (more negative V_G). In state 2' and 2, the electron is either (approximately) at the valence band or conduction band edge. The relative position of these potentials with respect to E_V does not change with bias.

- Occurrence of transient RTN, which is also likely due to a metastable state, similar to what has been suggested for anomalous RTN [28].

A particularly intriguing observation is that a defect can behave as a switching trap or not, depending on the drain bias. Considering that the capture time constants of all defects appear similar enough, it is thus most tempting to assume that all defects are basically of the same type and that the differences observed in $\bar{\tau}_e$ are merely due to small configurational differences. It should thus be obvious that a model that can describe such a wealth of information has to be more complicated than standard models and has to include metastable defect states.

In the following we will extend our previously suggested semiempirical model [10] which is based on the Harry-Diamonds-Lab switching trap model [32] for the E' center. As the exact microscopic nature of the defect responsible for recoverable NBTI has not been unanimously identified, it is worth highlighting that the E' center serves mainly as an inspiriation for the model but that the model can be applied to any defect which has the required metastable states. The configuration coordinate (CC) diagram of the extended version is shown in Fig. 10 and considers four different configurations of the defect, as have for instance been used by Poindexter for the E' center [33]. Two of the states are electrically neutral (1 and 1') while two of them correspond to the singly positively charged state (2' and 2). In each charge state the defect is represented by a double well, with the first of the two states being the equilibrium state and the other a secondary (metastable) minimum. Transitions involving charge exchange with the substrate are assumed to occur predominantly between 1 and 2' as well as 2 and 1'. On the other hand, transitions between 1 and 1' as well as between 2' and 2 are assumed to be purely thermally activated. In our previous model [10] the metastable state 1' had already been introduced in order to capture the bias dependence observed in the macroscopic measurements, in particular the switching trap behavior. While the fundamental assumptions employed in the old model appear to be basically correct, the rate equations where formulated in an *ad hoc* manner and the metastable state of defect 2 (2') was neglected. However, as will be shown below, to fully capture the non-linearity in the experimental data, state 2' has to be considered and the rate equations have to be derived more rigorously using nonradiative multiphonon (NMP) theory [34–36].

We assume that the time dynamics of the defect can be described by a simple stochastic process $X(t)$, a homogeneous continuous-time jump Markov process [37], see Fig. 11 (top). The defect can be in one of its four states i, described by the probabilities $p_i(t) = P\{X(t) = i\}$, with $i = 1, 1', 2, 2'$. Naturally, $\sum p_i(t) = 1$ at all times. Following conventional Markov process theory, the transition probabilities from state i to state j can be written as

$$P\{X(t+h) = j \,|\, X(t) = i; i \neq j\} = k_{ij}h + O(h), \tag{3}$$

with h being an infinitesimally small time step, $\lim_{h\to 0} O(h)/h = 0$, and k_{ij} as the transition rate. Together with the probability that the defect stays in state i

$$P\{X(t+h) = i \,|\, X(t) = i\} = 1 - h \sum_{j,i\neq j} k_{ij} + O(h), \tag{4}$$

it is straight forward to derive the partial differential equation (the master equation) controling the time evolution of the probabilities p_i

$$\frac{\partial p_i(t)}{\partial t} = \sum_j \Big(p_j(t) k_{ji} - p_i(t) k_{ij} \Big). \tag{5}$$

We neglect the transitions which cause a change in both the charge state and a considerable change in the defect configuration at the same time, that is, $k_{12} = k_{21} = k_{2'1'} = k_{1'2'} = 0$. Being a linear equation in $p_i(t)$, the expectation values of the probabilities, $E\{p_i(t)\} = f_i(t)$, are controlled by the same equation as (5), with f_i replacing p_i, which is then the conventional reaction rate equation. In other words, the stochastic description and the reaction rate formalism are 'consistent in the mean' for the linear problem at hand [38]. The advantage of using the stochastic description (5) lies in the fact that it also allows the stochastic simulation of RTN and tRTN as well as a simple calculation of the probability density function (p.d.f.) describing the stochastic capture and emission processes [39], see Fig. 11. With the p.d.f. at hand, the effective capture and emission times as well as their variances can be calculated analytically and compared to experimental data.

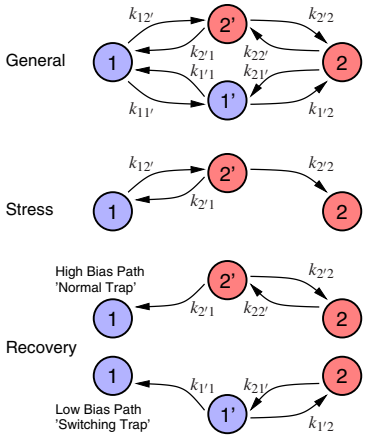

Fig. 11: State transition rate diagram for the general case (top), stress case (middle), and the two important pathways for recovery (bottom). During stress and recovery the Markov chain is described by a Birth-Death process since we are only interested in the first passage time from the initial to the final state, the expectation value of which gives the time constant.

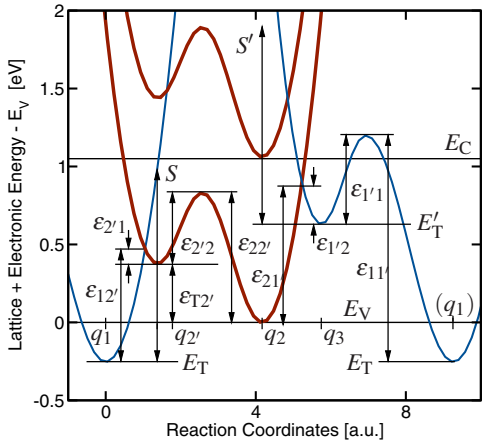

Fig. 12: Definition of the symbols used in the model. Note that the reaction coordinate describing the transition $1 \leftrightarrow 2'$ is different from the one describing $2 \leftrightarrow 1'$, which is why the potential describing states 1 and 1' is plotted twice, once to the left and once to the right of state 2. In this simple model only the expansions around q_1 and $q_{2'}$ determine $1 \leftrightarrow 2'$ while the expansions around q_2 and q_3 determine $2 \leftrightarrow 1'$.

The details of the derivation of the transition rates k_{ij} will be published elsewhere and we restrict ourselves to summarizing the main assumptions together with the results. Most assumptions employed in the model are standard and frequently used in NMP theory with a few important exceptions:

- We assume that the electron-phonon system can be dealt with in the Born-Oppenheimer approximation and describe the defect energy using adiabatic potentials.
- As the model is based on reported properties of the E' center, which has been suggested to undergo severe structural relaxation upon charge capture, we restrict our derivation to the strong electron-phonon coupling case.
- Since our prime concern is charge trapping at higher temperatures, such as observed during NBTS, the final equations can be considerably simplified (semiclassical approximation).
- To be somewhat more general than the frequently used theories, we consider both linear and quadratic electron-phonon coupling modes [40].
- In order to highlight the important features of the model, the transition rates are only given for the case of charge exchange with the substrate valence band. Charge exchange with the conduction band, which is important in accumulation (PBTI), as well as charge exchange with the gate can be dealt with analogously.
- Quite frequently one finds that the NMP equations derived for bulk traps are used for oxide defects, that is, the NMP process merely introduces a bias-independent thermal barrier ΔE_B [15, 41]. Our previously suggested model [10, 20] suffered from the same problem and in order to properly account for the experimentally observed bias dependence we empirically introduced a field assisted acceleration term, $\exp(-F^2/F_c^2)$, valid for the bulk case only [42, 43]. In fact, however, in oxide traps the relative position of the parabolas changes with bias [31, 36, 44], quite naturally introducing the required strong bias dependence into ΔE_B, in addition to some other features.

Under the above assumptions the relations for the transition rates involving charge transfer are

$$k_{12'} = \sigma v_{th} p e^{-\beta\varepsilon_{12'}}, \qquad k_{2'1} = \sigma v_{th} p e^{-\beta\varepsilon_{12'}} e^{-\beta(E_{TF}-\varepsilon_{T2'})}, \tag{6}$$

$$k_{1'2} = \sigma v_{th} p e^{-\beta\varepsilon_{1'2}}, \qquad k_{21'} = \sigma v_{th} p e^{-\beta\varepsilon_{1'2}} e^{-\beta E_{T'F}}. \tag{7}$$

For simplicity we assume all capture cross sections to be equal and given by a simple WKB approximation for a large and thin barrier of height ϕ as $\sigma = \sigma_0 \exp(-x/x_0)$, with $x_0 = \hbar/(2\sqrt{2m\phi})$ and v_{th} the thermal velocity.

For the calculation of the barriers $\varepsilon_{12'}$ and $\varepsilon_{1'2}$ the adiabatic potentials in each state i are expanded quadratically around the minima q_i, which in general results in different vibrational frequencies ω_i. Note that for linear electron-phonon coupling all frequencies are the same, $\omega_i = \omega$ [40]. Then, by setting $\omega_i = R_i \omega_j$ the rates can be approximately given as

$$\varepsilon_{ij} \approx \frac{S_i\hbar\omega_i}{(1+R_i)^2} + \frac{R_i}{1+R_i}\varepsilon_{Ti} + \frac{R_i}{4S_i\hbar\omega_i}\varepsilon_{Ti}^2, \tag{8}$$

where S_i is the Huang-Rhys factor characterizing the number of phonons required for the optical transition. The trap depth ε_{Ti} is either $\varepsilon_{T1} = E_V - E_T + \varepsilon_{T2'}$ or $\varepsilon_{T1'} = E_V - E'_T$. We remark that the above equation is exact for linear electron-phonon coupling ($R_i = 1$), and gives the familiar result $\varepsilon_{ij} = (\varepsilon_{Ti} + S_i\hbar\omega)^2/(4S_i\hbar\omega)$.

For simplicity, we assume the transition between states 1 and $1'$ as well as $2'$ and 2 to be bias independent but to occur along different reaction coordinates. Consequently, we do not calculate the barriers via intersections of the parabolas but consider them as explicit parameters. Obviously, $\varepsilon_{22'} = \varepsilon_{T2'} + \varepsilon_{2'2}$ and $\varepsilon_{11'} = \varepsilon_{1'1} + E'_T - E_T$ and we use $k_{mn} = \nu_m \exp(-\beta\varepsilon_{mn})$ where $\nu_m \sim 10^{13}\,\mathrm{s}^{-1}$.

Naturally, the prediction of the model depends heavily on an accurate description of the defect potentials which are described by the parameters $S\hbar\omega = S_1\hbar\omega_1$, $S'\hbar\omega' = S_{1'}\hbar\omega_{1'}$, $\Delta E_T = E_{T_0} - E_{V_0}$, $\Delta E'_T = E'_{T_0} - E_{V_0}$, $\varepsilon_{22'}$, $\varepsilon_{1'1}$, $R = R_1$, and $R' = R_{1'}$. In addition, the trap depth x as well as the capture cross section ($\sim 10^{-15}\,\mathrm{cm}^{-2}$) and the attempt frequency ($\sim 10^{13}\,\mathrm{s}^{-1}$) are required. While it appears that the model has more parameters than one might wish it to have, one has to recall that this is a consequence of the multitude of features it can explain, including the full bias and temperature dependence. Should one aim at a crude, first-order approximation, one could neglect the metastable states (no tRTN, no switching trap behavior, linear bias dependence of $\log(\bar{\tau}_c)$) to obtain a rough model of the defect with the three parameters $S\hbar\omega$, ΔE_T, x only. Care has to be taken, though, that the physical meaning of these effective parameters may then be questionable.

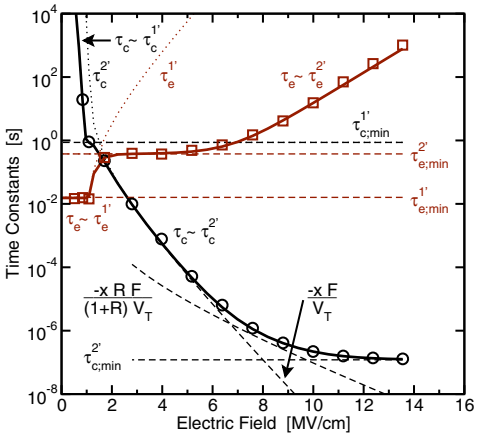

Fig. 13: Bias dependence of the effective capture and emission times according to the model. The symbols are obtained from a Monte Carlo simulation of the stochastic process, while the dashed lines give the individual contributions to the effective time constants (solid lines).

V. Approximate Solutions

Although the solution of the master equation (5) is in principle straight forward to obtain, it does not provide significant insight into the behavior of the defect. In particular, depending on the defect configuration various complicated transition patterns are possible, most notably patterns which would be recognized as RTN and anomalous RTN. However, during both stress and recovery the rates become highly asymmetric, strongly favoring a transition to 2 during stress and back to 1 during recovery, see Fig. 11. The most likely path during stress is from 1 to 2, while during recovery the defect may either recover via 2' or 1', the latter becoming particularly important at low $|V_G|$. Thus, in the following approximate solutions for these two important limiting cases will be given, while the details of the derivation will be published elsewhere. We only consider the dominant charge exchange with the valence band in the substrate. Also, the second-order term in ε_{Ti} defined by (8) will be neglected and the oxide will be assumed to be charge-free, in order to provide a simple relationship between the electrostatic defect level and the (constant) field in the oxide. Finally, Boltzmann statistics will be assumed which is valid for $\bar{\tau}_e$ at low $|V_G|$ but rather crude for $\bar{\tau}_c$ at high $|V_G|$. Nonetheless, the qualitative features of the model are not affected by either of these approximations.

The capture time constant is calculated from the expectation value of the first passage time from state 1 to state 2, see Fig. 11. The transition may proceed either via state 2', which is the dominant case at high bias, or via state 1' at lower biases (not shown in Fig. 11). For the first case we obtain $\bar{\tau}_c^{2'}$, while the latter case is described by $\bar{\tau}_c^{1'}$, which, in total, results in

$$1/\bar{\tau}_c \approx 1/\bar{\tau}_c^{2'} + 1/\bar{\tau}_c^{1'}. \tag{9}$$

Conversely, the emission time constant is calculated from the expectation value of the first passage time from state 2 to state 1 which may also proceed either via state 2' or via state 1'. For the first case we obtain $\bar{\tau}_e^{2'}$, while the latter case is described by $\bar{\tau}_e^{1'}$, which, again, gives

$$1/\bar{\tau}_e \approx 1/\bar{\tau}_e^{2'} + 1/\bar{\tau}_e^{1'}. \tag{10}$$

It is worth emphasizing that the capture and emission times described in such a way are not exponentially distributed. Nonetheless, it can be shown that under most bias conditions the time constants are dominated by a single transition, in which case the p.d.f. becomes nearly exponential, thereby justifying the concept of an effective first-order process with effective capture and emission times (9) and (10).

The individual contributions to the time constants are

$$\bar{\tau}_c^{2'} = \bar{\tau}_{c;\min}^{2'}\Big(1+\frac{N_1}{p}\exp\Big(-\frac{xF}{V_T}\Big)\Big) + \tau_0\frac{N_2}{p}\exp\Big(-\frac{xR}{1+R}\frac{F}{V_T}\Big), \tag{11}$$

$$\bar{\tau}_c^{1'} = \bar{\tau}_{c;\min}^{1'} + \tau_0\frac{N_3}{p}\exp\Big(-\frac{xR'}{1+R'}\frac{F}{V_T}\Big), \tag{12}$$

$$\bar{\tau}_e^{2'} = \bar{\tau}_{e;\min}^{2'} + \tau_{2'}\exp\Big(\frac{x}{1+R}\frac{F}{V_T}\Big), \tag{13}$$

$$\bar{\tau}_e^{1'} = \bar{\tau}_{e;\min}^{1'}\Big(1+e^{\beta E_{T'F}}\Big) + \tau_{1'}\exp\Big(\frac{x}{1+R'}\frac{F}{V_T}\Big). \tag{14}$$

with the same τ_0 as used in (2) and the temperature-dependent but field-independent auxiliary quantities

$$N_1 = N_V\exp(\beta(\varepsilon_{T2'} - \Delta E_T)),$$

$$N_2 = N_V\exp\Big(\beta\Big(\frac{S\hbar\omega}{(1+R)^2} - \frac{R(\Delta E_T - \varepsilon_{T2'})}{1+R}\Big)\Big),$$

$$N_3 = N_V\exp\Big(\beta\Big(\frac{S'\hbar\omega'}{(1+R')^2} - \frac{R'\Delta E_T'}{1+R'}\Big)\Big)(1+\exp(\beta(\Delta E_T' - \Delta E_T))),$$

$$\tau_{2'} = \tau_0\exp\Big(\beta\Big(\frac{S\hbar\omega}{(1+R)^2} + \frac{\Delta E_T - \varepsilon_{T2'}}{1+R}\Big)\Big)(1+\exp(\beta\varepsilon_{T2'})),$$

$$\tau_{1'} = \tau_0\exp\Big(\beta\Big(\frac{S'\hbar\omega'}{(1+R')^2} - \frac{\Delta E_T'}{1+R'}\Big)\Big).$$

Interestingly, since each time constant contains a contribution from a purely thermal transition, the minimum value is bounded by $\bar{\tau}_{c;\min}^{2'} = 1/k_{2'2}$, $\bar{\tau}_{c;\min}^{1'} = 1/k_{13}$, $\bar{\tau}_{e;\min}^{2'} = 1/k_{22'}$, and $\bar{\tau}_{e;\min}^{1'} = 1/k_{31}$. An evaluation of the analytic capture and emission time models against a Monte Carlo simulation of the full model is given in Fig. 13 for a switching trap, where around $V_G = V_{th}$ the dominant pathway changes from via 2' to 1'.

A. Normal Kinetics

Under 'normal' kinetics we understand the case where the impact of the metastable states is not directly obvious, that is, no switching behavior can be observed (no transition to 1'). This is the case when $\Delta E_T'$ is too large to give a significant occupancy of state 1'. In this case, the time constants are

$$\bar{\tau}_c = \bar{\tau}_{c;\min}^{2'}\Big(1+\frac{N_1}{p}\exp\Big(-\frac{xF}{V_T}\Big)\Big) + \tau_0\frac{N_2}{p}\exp\Big(-\frac{xR}{1+R}\frac{F}{V_T}\Big), \tag{15}$$

$$\bar{\tau}_e = \bar{\tau}_{e;\min}^{2'} + \tau_{2'}\exp\Big(\frac{x}{1+R}\frac{F}{V_T}\Big). \tag{16}$$

Both time constants consist now of two terms, where the first one denotes the impact of the relaxation barrier $\varepsilon_{2'2}$. For capture, the field dependence of the two terms is different, resulting in a non-linear overall exponential field dependence and eventual saturation at $\bar{\tau}_c = \bar{\tau}_{c;\min}^{2'}$ for high fields. For example, for $R = 1$, the argument of the exponential field term is initially $-xF/V_T$ and gradually reduces to $-xF/2V_T$ with increasing field. In addition, a weaker bias dependence is introduced by the same $1/p$ dependence of both factors. Note that this $1/p$ dependence is the only field dependence of $\bar{\tau}_c$ in the standard SRH-like model (2).

For emission, the term due to the relaxation barrier is bias independent and dominates for small fields. This is similar to (2) where irrespective of the exponential F dependence only a weak field dependence is obtained, as F depends weakly on V_G below V_{th}. For $R = 1$ one obtains $xF/2V_T$ for the argument of the exponential field term at high fields. In standard SRH-like models the argument is twice as large, namely xF/V_T. Note how NMP theory with a linear

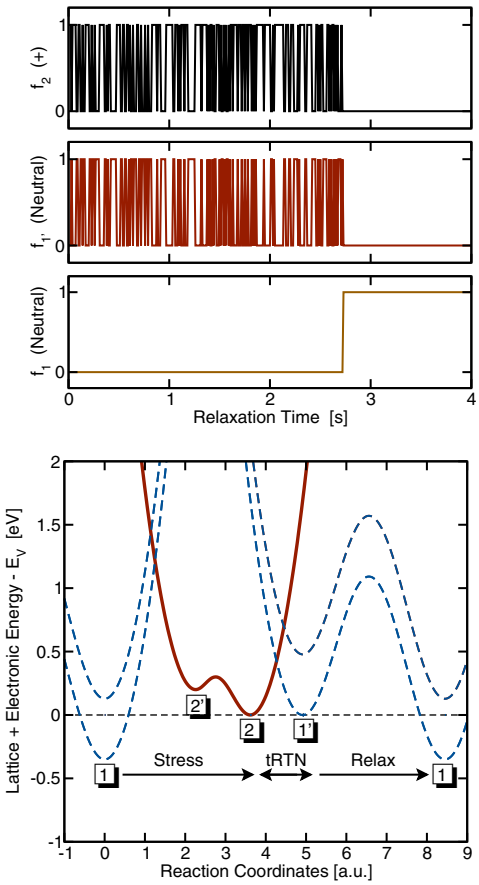

Fig. 14: Stochastic simulation (top) of a defect configuration (bottom) that leads to tRTN. During stress, state 2 is occupied, during the initial recovery phase the defect switches between states 2 and 1', while eventually the defect anneals. Such a behavior can only be observed in a very narrow window of V_G where the minima of states 2 and 1' are close to each other.

coupling term only ($R = 1$) introduces a symmetric field dependence in both $\bar{\tau}_c$ and $\bar{\tau}_e$ which becomes asymmetric for $R \neq 1$. Nonetheless, for medium values of F the ratio $\bar{\tau}_c/\bar{\tau}_e$ is controlled by $-xF/V_T$, just as in the SRH-like model (2).

B. Switching Trap Kinetics

Under the condition that the metastable state 1' is moved close to E_V and the barrier separating 2 and 1' is low enough, a transition $2 \rightarrow 1'$ can occur and the defect may even switch back and forth between states 1' and 2, see Fig. 11 (bottom). For normal switching traps, however, these transitions are too fast to be directly observable by the measurement equipment which only records the average value.

In the switching trap configuration, the impact of the metastable state 1' becomes evident in $\bar{\tau}_e$, see Fig. 13. At low enough bias, annealing of the defect back to state 1 will now occur via state 1'. Note that although even during stress the pathway $1 \rightarrow 1' \rightarrow 2$ is theoretically possible, we have so far not observed a defect compatible with such a configuration.

Introducing $f_n = (1+k_{1'2}/k_{2'1})^{-1} = (1+\exp(\beta E_{T'F}))^{-1}$, which is the probability that the trap level E'_T is occupied by an electron, that is, neutral, we can express the emission time constant for $\bar{\tau}_e^{1'} \lesssim \bar{\tau}_e^{2'}$ as

$$\bar{\tau}_e = \frac{\bar{\tau}_{e;min}^{1'}}{f_n} + \tau_{1'} \exp\Big(\frac{x}{1+R'}\frac{F}{V_T}\Big). \qquad (17)$$

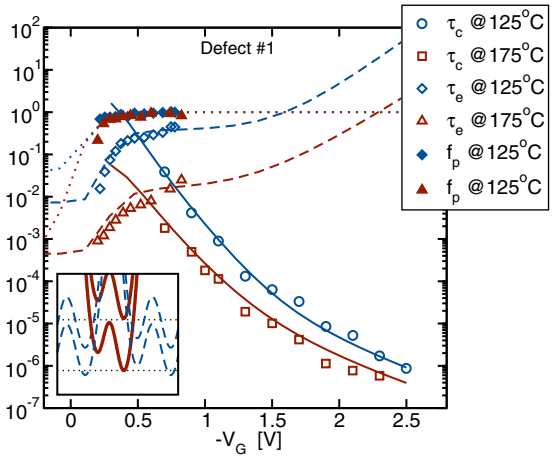

Fig. 15: Simulated capture and emission time constants for defect #1 compared to the experimental values. The CC diagram shown in the inset is similar to that of the tRTN case shown in Fig. 14 with the difference that the barrier between states 2 and 1' is rather small. The experimental occupation probability f_p is given by the filled symbols.

This is a remarkable result: when $E_F > E'_T$, the defect is neutral ($f_n = 1$) and the emission time is given by the bias independent value $\bar{\tau}_{e;min}^{1'}$. As soon as the defect level E'_T moves above the Fermi level, the probability f_n will decrease, thereby strongly increasing $\bar{\tau}_e$. This strong bias dependence explains the typical switching trap characteristics around the threshold voltage observed experimentally in Fig. 7. For large $|V_G|$, on the other hand, $\bar{\tau}_e^{1'}$ will become larger than $\bar{\tau}_e^{2'}$ and the pathway $2 \rightarrow 2' \rightarrow 1$ dominates the emission time.

An interesting special configuration of the switching trap leads to tRTN, namely when the transitions between states 1' and 2 are slow enough to fall within the experimental window, as discussed in Section VI.

VI. Qualitative Model Behavior

A stochastic simulation of a defect configuration leading to tRTN is shown in Fig. 14. During stress, state 2 becomes occupied ($p_2 = 1$), while during the initial recovery phase the defect switches back and forth between states 2 and 1' ($p_2 + p_{1'} = 1$), visible as tRTN. Eventually, the defect anneals by a transition from 1' to 1 ($p_1 = 1$). Such a behavior can only be observed in a very narrow window of V_G where the minima of the states 2 and 1' are close to each other. Furthermore, the barrier between states 2 and 1' must be large enough to cause capture and emission times within the experimental window. For transitions faster than experimentally observable, the measurement equipment will record the averaged signal $E\{P\{X(t) = 2\,|\,X(t) = 2 \vee 1'\}\} = f_p = 1 - f_n$ and the defect will appear as a switching trap.

VII. Quantitative Model Behavior

The simulated capture and emission time constants calibrated to the experimental data available for defect #1 are shown in Fig. 15. As can be seen from the CC diagram shown in the inset, defect #1 is similar to the schematic tRTN case shown in Fig. 14 with the difference that the barrier between states 2 and 1' is rather small. As a consequence, the fluctuations between states 2 and 1' are too fast and cannot be resolved by the measurement equipment and the defect appears like a switching trap.

In contrast, Fig. 16 shows the model calibrated to the data of defect #4. Here the metastable state 1' is energetically too high and has thus no apparent effect on the measured capture and emission times.

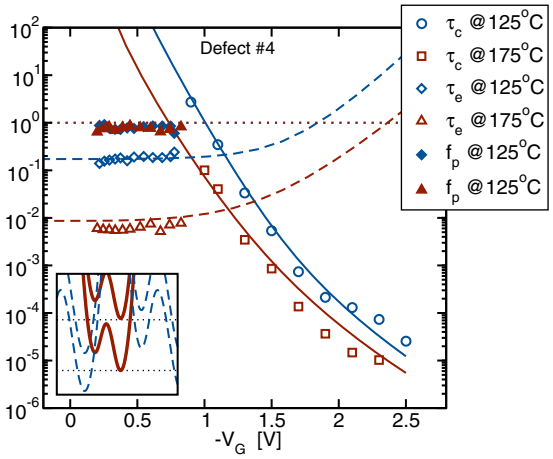

Fig. 16: Similar to Fig. 15, but now for defect #4. In contrast to defect #1, E'_T is too high to be significantly populated and no switching trap behavior can be observed. This can be clearly seen in the CC diagram shown in the inset. As a consequence, the emission time constant is insensitive to the gate bias.

As can be seen from these examples, the suggested model can naturally explain all experimentally observed features. In particular the initially surprising occurrence of tRTN, switching traps, as well as the non-linearities in $\log(\bar{\tau}_c)$ and $\log(\bar{\tau}_e)$ have been explained consistently within a single model.

VIII. Conclusions

We have suggested the powerful time dependent defect spectroscopy (TDDS) for the analysis of defects leading to NBTI and RTN. Application of this new method in a large number of experiments leads us to the following conclusions: *(i)* NBTI recovery is due to discharging of individual defects with a wide distribution of time scales. There is no diffusion involved. *(ii)* Both capture and emission time constants are temperature activated, consistent with nonradiative multiphonon theory. Furthermore, the time constants are uncorrelated. No signs of a temperature-independent elastic tunneling process could be found. *(iii)* The capture time constants show a very strong field dependence. Similarly, the bias dependence of the emission time constant around V_{th} may be either weak or strong, depending on the configuration of the defect. In addition, the visibility of this switching trap behavior my depend on the bias conditions. *(iv)* The total number of defect precursors is pre-existing and no signs of newly created defects which create emission events within our experimental window could be found in our short to medium term stress experiments. *(v)* The existence of metastable states becomes obvious due to disappearing defects, transient RTN, and a bias dependencies stronger than expected in a simpler model. *(vi)* Finally, the defects responsible for the recoverable component of NBTI are identical to those causing RTN. Conceptually, the difference is as follows: in an RTN experiment only a limited number of defects having capture and emission time constants within the experimental window are visible. Due to the strong bias dependence of the capture time constant, many more defects contribute to NBTI. These defects can be visualized in a TDDS setup. Consequently, NBTI can be considered the non-equilibrium response of these defects while RTN is a consequence of their quasi-stationary behavior.

We have finally suggested a comprehensive defect model which can describe the standard and switching trap defect behavior as well as all the anomalies observed, such as disappearing defects and transient RTN. Although the model is based on reported properties of the E' center, in particular the switching behavior, we do at the moment not have sufficient evidence to claim that the NBTI recoverable component is solely (or at all) due to E' centers, particularly in nitrided and high-k oxides. Nonetheless, as the model uses empirical potentials which are adjusted to fit the experimental data, it is perfectly general and can be applied to a broad range of defects having metastable states. The model can accurately describe the bias dependence of the capture time constant which thus allows back-extrapolation from stress to operating conditions. Furthermore, the model accurately describes the bias dependence of the emission time constant around V_{th}, which is important for our understanding of the functioning of a transistor in a circuit where the gate bias is not just switched between static stress and recovery voltages.

IX. 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 Th. Aichinger, Ph. Hehenberger, Ph. Roussel, M. Uren and M. Kirton.

References

[1] T. Grasser *et al.*, in *Defects in Microelectronic Materials and Devices*, edited by D. Fleetwood *et al.* (Taylor and Francis/CRC Press, 2008), pp. 1–30.
[2] K. Jeppson *et al.*, JAP **48**, 2004 (1977).
[3] M. Alam *et al.*, MR **47**, 853 (2007).
[4] S. Mahapatra *et al.*, T-ED **56**, 236 (2009).
[5] S. Rangan *et al.*, *IEDM* (2003), pp. 341–344.
[6] B. Kaczer *et al.*, APL **86**, 1 (2005).
[7] V. Huard *et al.*, MR **46**, 1 (2006).
[8] H. Reisinger *et al.*, T-DMR **7**, 119 (2007).
[9] B. Kaczer *et al.*, *IRPS* (2009), pp. 55–60.
[10] T. Grasser *et al.*, *IRPS* (2009), pp. 33–44.
[11] T. Tewksbury, Ph.D. Thesis, MIT, 1992.
[12] B. Kaczer *et al.*, *IRPS* (2005), pp. 381–387.
[13] A. McWhorter, Sem.Surf.Phys 207 (1957).
[14] M. Weissman, Rev.Mod.Phys **60**, 537 (1988).
[15] M. Kirton *et al.*, Adv.Phys. **38**, 367 (1989).
[16] V. Huard *et al.*, *IIRW* (2005), pp. 5–9.
[17] B. Kaczer *et al.*, *IRPS* (2008), pp. 20–27.
[18] H. Reisinger *et al.*, *IRPS* (2010), (in print).
[19] H. Reisinger *et al.*, *IIRW* (2009), pp. 30–35.
[20] T. Grasser *et al.*, *IEDM* (2009), pp. 729–732.
[21] D. Lang, JAP **45**, 3023 (1974).
[22] A. Karwath *et al.*, APL **52**, 634 (1988).
[23] D. Vuillaume *et al.*, PRB **34**, 1171 (1986).
[24] A. Asenov *et al.*, T-ED **50**, 839 (2003).
[25] H. Mueller *et al.*, JAP **79**, 4178 (1996).
[26] B. Kaczer *et al.*, *IRPS* (2010), (in print).
[27] D. Gillespie, J.Comp.Phys. **22**, 403 (1976).
[28] M. Uren *et al.*, PRB **37**, 8346 (1988).
[29] T. Grasser *et al.*, *IEDM* (2007), pp. 801–804.
[30] K. Ralls *et al.*, PRL **52**, 228 (1984).
[31] A. Avellán *et al.*, JAP **94**, (2003).
[32] A. Lelis *et al.*, T-NS **41**, 1835 (1994).
[33] E. Poindexter *et al.*, J.Electrochem.Soc. **142**, 2508 (1995).
[34] K. Huang *et al.*, Proc.R.Soc.A **204**, 406 (1950).
[35] C. Henry *et al.*, PRB **15**, 989 (1977).
[36] W. Fowler *et al.*, PRB **41**, 8313 (1990).
[37] D. Gillespie, *Markov Processes* (AP, 1992).
[38] D. McQuarrie, J.Appl.Prob. **4**, 413 (1967).
[39] O. Ibe, *Markov Processes for Stochastic Modeling* (AP, 2009).
[40] C. Kelley, PRB **20**, 5084 (1979).
[41] P. Restle *et al.*, IBM J.Res.Dev. **34**, 227 (1990).
[42] S. Makram-Ebeid *et al.*, PRB **25**, 6406 (1982).
[43] S. Ganichev *et al.*, Phys.SS **39**, 1703 (1997).
[44] A. Palma *et al.*, PRB **56**, 9565 (1997).