Microelectronics Reliability 52 (2012) 39–70

Contents lists available at SciVerse ScienceDirect

Microelectronics Reliability

journal homepage: www.elsevier.com/locate/microrel

Introductory Invited Paper

Stochastic charge trapping in oxides: From random telegraph noise to bias temperature instabilities

Tibor Grasser

Institute for Microelectronics, TU Wien, Gusshausstrasse 27-29, 1040 Wien, Austria

ARTICLE INFO

Article history:
Received 4 September 2011
Accepted 5 September 2011
Available online 2 October 2011

ABSTRACT

Charge trapping at oxide defects fundamentally affects the reliability of MOS transistors. In particular, charge trapping has long been made responsible for random telegraph and $1/f$ noise. Recently, it has been identified as a significant contributor to bias temperature instabilities. Conventional defect models assume that the defect has two states, one of them neutral and the other charged. The transition rates between the two states are calculated using some extended Shockley–Read–Hall theory, which neglects the configurational changes occurring at the defect site following a charge trapping or emission event. In order to capture these changes, multiphonon models have been in use for many decades but have not found their way into the mainstream of reliability modeling yet. Furthermore, recent experimental results demonstrate that defects have more states than the two assumed in the conventional model. These additional states together with multiphonon charge transfer mechanisms are essential for the understanding of the complex defect dynamics. The present review summarizes the basic principles of how to model stochastic defect transitions with a particular focus on multi-state defects. After discussing the limitations of Shockley–Read–Hall theory, the relatively simple semiclassical approximation of multiphonon theory is introduced which already provides a much better description. Finally, the transition rates for multi-state defects are estimated using multiphonon theory, which gives a very accurate description of the latest experimental data.

1. Introduction

The non-ideal behavior of metal–oxide–semiconductor field effect transistors (MOSFETs) is essentially determined by defects at the semiconductor–insulator interface as well as inside the insulating oxide. Since their detailed microscopic nature is still controversial, a phenomenological classification into interface states (fast states), border states (slow states), and oxide states (fixed oxide charge, anomalous positive charge, etc.) is often employed [1,2]. While the fast interface states are commonly attributed to P_{b0} and P_{b1} centers, which are trivalent silicon dangling bonds at the SiO_2/Si interface, opinion is divided on border states. Similarly to interface states, these border states also communicate with the underlying silicon channel by exchanging charge carriers, albeit on larger time scales. In the simplest theory, the transition rates decrease exponentially with the distance of the defect from the interface. However, given the ultrathin dielectrics employed in modern silicon technology, this decrease is not significant enough to prevent defects from capturing a charge at any position inside the oxide, quite in contrast to technologies with thick oxides. As such, in modern CMOS transistors every defect inside the oxide must be considered a potential border state, particularly if also the interaction with the gate is considered.

Border states are often associated with E' centers (trivalent silicon dangling bonds in the oxide) [5,6], but have also been related to hydrogenic defects [7,8]. Border states are commonly considered the cause of random telegraph and $1/f$ noise [9]. In addition, due to their wide distribution of time constants, they have been suspected to cause slow drifts in crucial transistor parameters such as the threshold voltage in a phenomenon that has become known as the bias temperature instability (BTI) [10–14]. While in pMOSFETs the most relevant form is the negative bias temperature instability (NBTI), which is observed under larger negative gate voltages, in nMOSFETs employing high-k oxides the positive bias temperature instability (PBTI) is of high technological interest [15–17].

The most popularized explanation for BTI invokes an interfacial Si–H breakage process which for longer stress time is controlled by the diffusion of the released hydrogen inside the oxide [10,18–20]. This reaction–diffusion model has received a lot of criticism recently as it is unable to explain a variety of experimentally observed features [21–26,13,27–30]. In contrast, application of theoretical RTN models to BTI analysis already leads to a significant improvement in the model quality [31,13]. The idea behind this concept is schematically shown in Fig. 1: under stationary conditions, defects randomly exchange charge with the substrate,

E-mail address: grasser@iue.tuwien.ac.at

0026-2714/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.microrel.2011.09.002

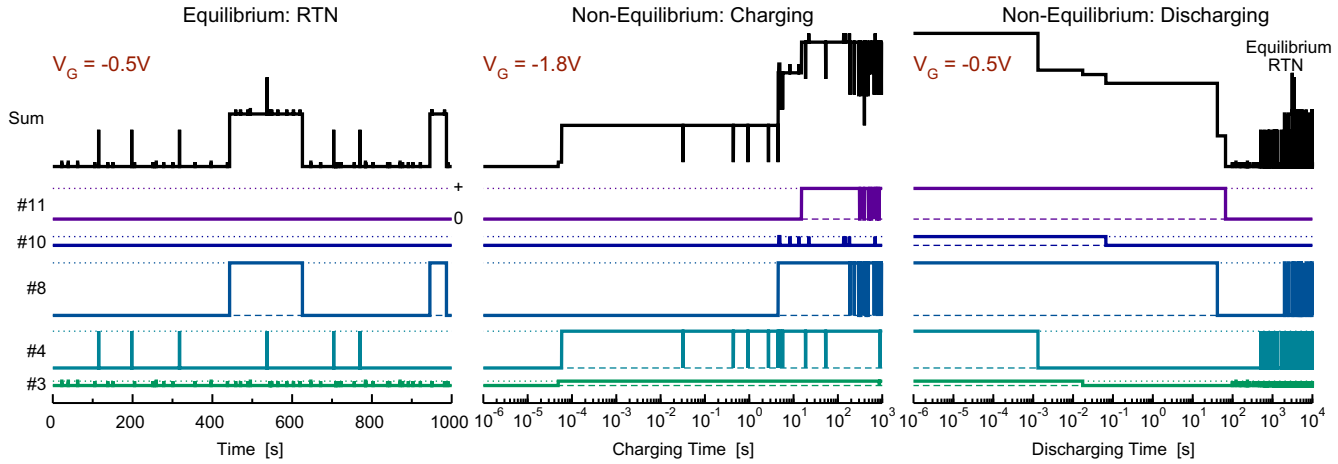

Fig. 1. Comparison of charge capture and emission events under stationary RTN conditions (left) with the synchronized capture and emission events during the BTI (middle and right). For clarity, only five of the thirteen defects of the example pMOS studied in Refs. [3,4] are used. Their stochastic behavior was simulated using real parameters extracted by the time-dependent defect spectroscopy (TDDS) at 175 °C. **Left**: Close to the threshold voltage ($V_G = -0.5$ V), the RTN is dominated by defect #3 with the occasional contribution from defect #4 and #8. Defects #10 and #11 have larger time constants (400 ks and 1.5 Ms) and remain neutral within the experimental window. **Middle**: Application of the charging voltage ($V_G = -1.8$ V) results in a non-equilibrium response of the defects which are assumed to be initially discharged. Due to the strong field dependence of $\bar{\tau}_c$, the defects become predominantly positively charged. For larger charging times ($t_s \gtrsim \bar{\tau}_c$), the RTN produced by each defect is visible on the logarithmic scale. **Right**: Following the perturbing charging step, defects with $\bar{\tau}_c \lesssim t_s$ are likely to be synchronized in their charged states. A switch back to a lower voltage again results in a non-equilibrium response until equilibrium is reached after the longest decorrelation time due to defect #11. Each discrete step in the transient is due to the emission of a single hole, following its emission time constant. At the end of the recovery trace, equilibrium RTN is observed again.

leading to RTN and $1/f$ noise. During stress, the bias is switched to a larger value and the capture time constants of the defects become much smaller due to their strong bias-dependence. As a consequence, defects with smaller capture than emission times will be preferentially in their charged states. Given the wide distribution of the time constants, these transitions to the charged states will occur at different times for each defect. Summing up the individual transitions results in what is conventionally observed as BTI degradation. When the gate bias is switched back to the pre-stress value, the defects return to their pre-stress occupancy, visible as the ubiquitous BTI relaxation transients [32]. In that context it is interesting to remark that a considerable number of publications exist which show that during bias temperature stress defects with longer recovery time constants are activated [21,33,34]. As such, these defects do not return to their pre-stress occupancy within a reasonable measurement time when the stress is removed and appear as a permanent component [35,21,22,33,36–39,34]. Unfortunately, the microscopic nature of this permanent component is still unclear [40,41,28,29,42,34]. Still, in typical experimental windows, starting from the microsecond regime up to weeks, both degradation as well as recovery seem to be dominated by border states [34].

From a theoretical point of view, the transition rates between the two defect states are conventionally modeled using a standard Shockley–Read–Hall model (SRH) [43,44]. While the fast interface states appear to be compatible with SRH theory, it is difficult to make more precise statements, since their response to external stimulus occurs faster than can be directly measured. For example, these states impact the subthreshold slope of MOS transistors, as their occupancy can quickly follow changes in the gate bias. Indirect evidence for the correctness of SRH theory is available in the agreement of experiment and theoretical prediction seen in charge-pumping measurements [45]. In contrast, the behavior of border states is more complicated. Initial modeling attempts tried to describe the wide distribution of capture and emission time constants of these border states in the spirit of the SRH model. The assumption was that SRH theory is essentially valid when the spatial separation of the defect and the channel is considered by an additional tunneling term [46]. However, it has long been understood that charge exchange between border states and the channel occurs via a multiphonon rather than an elastic tunneling process and that the large time constants are not primarily due to the spatial depth of the defect but rather due to its thermal barrier upon charge capture [47–50]. One reflection of this fact is that the field-dependence of both the capture as well as the emission time constant cannot be properly explained by a tunneling process through a barrier which is only a few electron-volts high and about a nanometer thick [48]. Similar considerations hold for the temperature activation of the time constants as well as the fact that experimental time constants are much larger than can be otherwise explained in ultrathin modern gate stacks [51].

Conventional defect models assume that the defect can exist in two states, one of them charged and the other neutral. For instance, in an RTN experiment the drain current would switch between two discrete current levels, with the transition times being exponentially distributed, consistent with a two-state Markov process [50], see for example Fig. 2. However, deviations from this simple first-order model have been observed occasionally: a particularly intriguing example was described by Uren et al. [52] who noticed that a conventional random telegraph signal appeared to be modulated by another, much slower signal. This modulation resulted in a complete disappearance of the RTN for statistically distributed times. Uren et al. concluded that this anomalous RTN could be best described by a defect having an additional metastable state. They estimated as a lower bound that about 4% of the defects would fall into this category. Naturally, this estimated percentage is likely to be too low [52], as for a defect to be observable as anomalous RTN, both the time-constants leading to RTN as well

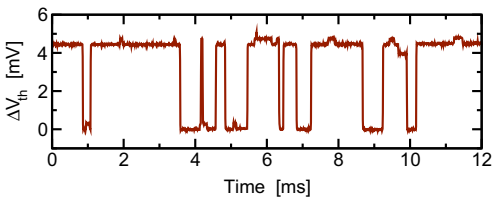

Fig. 2. Typical experimental RTN signal, recorded on a 2.2 nm pMOSFET biased around the threshold voltage.

as the time-constants leading to the disappearance and re-appearance of the defects must fall into the experimental window. Consequently, even if a defect falls produces anomalous RTN, it will likely appear as either a normal defect or even go completely unnoticed in a standard RTN experiment.

Recent experimental and theoretical results strongly suggest that such metastable defect states are the norm rather than the exception [3,4]. The impact of these states can be observed in various ways: First, inclusion of metastable transition states appears essential for an accurate description of the bias-dependence of the capture and emission time constants [4]. Second, they can create anomalous RTN as observed by Uren et al. [52] previously. Third, a stimulated variant of anomalous RTN has been recently observed following NBTI stress, called temporary RTN. And finally, defects can disappear and re-appear for stochastic amounts of time, also on very large timescales (months). All these results imply that transitions to or via metastable states occur on a wide range of timescale, fast for the bias-dependence, in the seconds regime for anomalous RTN, and in the very slow regime for disappearing defects. As such, the width of the distribution of time constants is comparable to the width of the distribution of the 'normal' capture and emission time constants and should therefore be an essential aspect of any defect model.

The following is an attempt at reviewing the present understanding of two- and multi-state defect dynamics. Starting with the conventional Markov theory for a two-state defect, the formalism is applied to multi-state defects. Then, experimental evidence is summarized, showing the bias- and temperature-dependence of the defects and some examples of 'anomalous' (multi-state) defect behavior. An essential part is the understanding of the capture and emission time constants, which do not follow SRH theory. In order to improve on the situation, multiphonon theory is summarized. Since from an application point of view the classical limit of multiphonon theory appears sufficient, the discussion is restricted to this limit, which results in rather compact and intuitive expressions. Finally, multiphonon theory is used to calculate the transition rates of multi-state defects, which eventually leads to excellent agreement with experimental data.

2. Stochastic defect modeling

Irrespective of the physical mechanism invoked to describe the actual capture and emission processes, most models assume that the defect can exist in two states, one neutral and the other charged. One speaks of donor-like defects when they can become positively charged and of acceptor-like defects when they can become negatively charged. Interface states, on the other hand, are typically amphotheric, meaning they have three states: positive, neutral, and negative. Furthermore, a defect may have metastable states. Metastable states are states which are not occupied under equilibrium conditions but impact the dynamic defect behavior, for instance during charging or discharging. Transitions to metastable states may occur without any change in the charge state. For instance, a defect may have an alternative (metastable) singly charged state which can be either reached from the neutral or from the 'normal' charged state. Transitions to the metastable state involve a different set of transition rates. As such, these metastable states result in a more complicated defect behavior compared to a simple two-state defect, which is sometimes referred to as anomalous [52].

In most cases, the stochastic transitions between the states can be described by a Markov process. Markov processes are widely used in many fields of science and a correspondingly wide range of literature is available [53,54]. The essence of a Markov process is that it is memory-less. In our context this means that the next transition depends solely on the current state, irrespective of how the defect got into the current state.

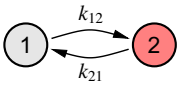

Fig. 3. State transition rate diagram for a simple two-state defect.

2.1. Two-state defects

We start our discussion with a simple two-state defect model. There, the defect has to be in one of its two states, which we shall call 1 and 2, see Fig. 3. For the sake of clarity lets assume that the defect is electrically neutral in state 1 and charged in state 2. Note that the actual charge state becomes relevant only when physical models for the transition rates are derived, see Section 5. The occupancies of each state are given by $X_i(t)$, with $X_i(t) = 1$ when the defect is in state i, and $X_i(t) = 0$ otherwise. In contrast to the occupation probability introduced below, only integer values 0 and 1 are allowed for defect occupancies, meaning that the defect has to be in a well-defined state at any time. Since the defect can only be in one of its two states, we have $X_1(t) + X_2(t) = 1$ at all times.

In order to derive the transition probabilities for a Markov process, we assume that the defect is in state i at time t. The probability for a transition to state j within the next infinitesimally small time interval h is given by the transition rate k_{ij}, which is a probability per unit time. For instance, the conditional probability that during the time interval h a transition from state 1 to state 2 occurs, given that the defect is already in state 1 at time t, is written as

$$P\{X_2(t+h) = 1 | X_1(t) = 1\} = k_{12}h + O(h), \tag{1}$$

with $\lim_{h\to 0} O(h)/h = 0$. Conversely, the probability that within h no transition from state 2 to state 1 occurs is given by

$$P\{X_2(t+h) = 1 | X_2(t) = 1\} = 1 - k_{21}h + O(h). \tag{2}$$

In the following we assume h to be so small that all higher-order terms represented by $O(h)$ are negligible. Introducing the shorthand $p_i(t) = P\{X_i(t) = 1\}$, the probability that $X_2(t+h)$ equals 1 can thus be written as

$$p_2(t+h) = P\{X_2(t+h) = 1 | X_1(t) = 1\} p_1(t), \tag{3}$$
$$P\{X_2(t+h) = 1 | X_2(t) = 1\} p_2(t), \tag{4}$$

since at time t the defect has to be in either of its two states. We can now replace the conditional probabilities in (4) by the rates (1) and (2) to obtain

$$p_2(t+h) = k_{12}hp_1(t) + (1 - k_{21}h)p_2(t)$$

which can be rearranged as

$$\frac{p_2(t+h) - p_2(t)}{h} = k_{12}p_1(t) - k_{21}p_2(t).$$

Expressing $p_1(t)$ by $1 - p_2(t)$ we obtain

$$\frac{dp_2(t)}{dt} = k_{12}(1 - p_2(t)) - k_{21}p_2(t), \tag{5}$$

in the limit of infinitesimally small h. This is an ordinary differential equation for $p_2(t)$ and has the solution

$$p_2(t) = p_2(\infty) + (p_2(0) - p_2(\infty))e^{-t/\tau}, \tag{6}$$

with

$$p_2(\infty) = \frac{k_{12}}{k_{12} + k_{21}} \quad \text{and} \quad \tau = \frac{1}{k_{12} + k_{21}}. \tag{7}$$

Eq. (6) describes the probability of the defect being in state 2 as an exponential transition from its initial value $p_2(0)$ to its final stationary value $p_2(\infty)$. Note that under stationary conditions $dp_2(t)/dt = 0$, meaning that the probability does no longer change with time. For instance, with $k_{12} = k_{21}$ the defect will have a 50% probability of being in either of its states. As such, the defect keeps hopping back and forth between its two states. Stationary conditions are reached for times much larger than the time constant τ, which is known as the asymptotic decorrelation time [53]. The meaning of this is as follows: at time $t = 0$ the defect is known to be in a state 2 with probability $p_2(0)$. A short time Δt later, with $\Delta t \ll \tau$, the probability $p_2(\Delta t)$ will still be close to $p_2(0)$. Only when $\Delta t > \tau$ will $p_2(\Delta t)$ be uncorrelated to the initial condition $p_2(0)$.

From (6) we can trivially calculate $p_1(t) = 1 - p_2(t)$.

$$p_1(t) = p_1(\infty) + (p_1(0) - p_1(\infty))e^{-t/\tau}, \tag{8}$$

with

$$p_1(\infty) = 1 - p_2(\infty) = \frac{k_{21}}{k_{12} + k_{21}}.$$

Together Eqs. (6) and (8) form what is known as a *Master equation*, which fully describes our simple two-state defect. The solution of the Master equation is shown Fig. 4 for a particular set of initial conditions.

2.2. Capture and emission time constants

Lets assume that the defect is in state 1 at $t = 0$. It is interesting to ask how long it will take for a transition to state 2 to occur. The time we have to wait for such a transition is known as the *first passage time* from state 1 to state 2, τ_{12}. Since we start off in state 1 and wait for the first transition to state 2 to occur, τ_{12} will be independent of the backward rate k_{21}. As a consequence, the defect model of Fig. 3 can be simplified to the one given in Fig. 5. With $p_1(0) = 1$ and $k_{21} = 0$, the Master equation is even simpler, yielding the result

$$p_1(t) = \exp(-k_{12}t). \tag{9}$$

The above can be used to calculate τ_{12}, which is the time point when the transition actually takes place. Obviously, τ_{12} is a stochastic variable, and we can calculate its probability density function (p.d.f.) by considering the following: the probability that the defect is already in state 2 at a certain time t is given by $p_2(t) = 1 - p_1(t)$. In such a case we know that τ_{12} must be smaller than t. As such we obtain the probability function of the random variable τ_{12} as

$$P\{\tau_{12} \leqslant t\} = p_2(\tau_{12}) = 1 - \exp(-k_{12}\tau_{12}). \tag{10}$$

Thus, the p.d.f. is

$$g(\tau_{12}) = \frac{dp_2(\tau_{12})}{d\tau_{12}} = k_{12}\exp(-k_{12}\tau_{12}), \tag{11}$$

meaning that τ_{12} is exponentially distributed with parameter $1/k_{12}$. Since in our example the defect has captured a hole in state 2, τ_{12} corresponds to the capture time τ_c from a physical point of view. We obtain the average capture time $\bar{\tau}_c$ as the expectation value of the exponential distribution, which is obviously given by

$$\bar{\tau}_c \triangleq \mathrm{E}\{\tau_c\} = \int_0^\infty \tau_c g(\tau_c)\,d\tau_c = \frac{1}{k_{12}}. \tag{12}$$

Similar arguments hold for the emission time, which describes the first transition from state 2 to state 1 provided the defect was in state 2 at $t = 0$, and one obtains $\bar{\tau}_e = 1/k_{21}$.

During the analysis of RTN, the times the defect spends in either of its states is usually collected into histograms. By normalizing these two histograms, exponential p.d.f.s are obtained which allow for the extraction of $\bar{\tau}_c$ and $\bar{\tau}_e$ [51]. Quite to the contrary, in the recently suggested time-dependent defect spectroscopy (TDDS) [3], BTI recovery traces are recorded on a logarithmic time scale, see Fig. 1. As a much larger number of defects with widely different emission times contribute to BTI, the use of a logarithmic scale is required. When an exponentially distributed quantity is binned into equidistant bins on a *logarithmic* axis, the shape of the p.d.f. changes, as this corresponds to a transformation of the random variable. So rather than considering the p.d.f. of τ, we have to consider the p.d.f. of $\theta = \log(\tau/\tau_0)$, with τ_0 being a normalizing constant. Starting from the definition of the expectation value for an arbitrary function $h(\tau)$, we obtain via a conventional variable transformation from τ to θ

$$\begin{aligned}\mathrm{E}\{h(\tau)\} &= \int_0^\infty h(\tau)g(\tau)\,d\tau = \int_{-\infty}^\infty h(\tau(\theta))g(\tau(\theta))\tau(\theta)\,d\theta \\ &= \int_{-\infty}^\infty h(\tau(\theta))\tilde{g}(\tau(\theta))\,d\theta,\end{aligned} \tag{13}$$

where

$$\tilde{g}(\tau) = \tau g(\tau) \tag{14}$$

is the p.d.f. on a logarithmic scale. The corresponding logarithmic p.d.f. of the exponential distribution (11) is thus

$$\tilde{g}(\tau) = \frac{\tau}{\bar{\tau}}\exp\left(-\frac{\tau}{\bar{\tau}}\right), \tag{15}$$

and is shown Fig. 6. We will meet this distribution again during the analysis of the spectral maps obtained by the TDDS, see Section 3.2.

2.3. Moments of the probability distribution

Stochastic processes are often characterized by their moments, for instance the expectation value and variance. These moments are straight-forward to calculate once the solution of the Master equation is available. For instance, consider the occupancy of state 2, X_2. We have seen previously that the probability of being in state 2 is p_2, where we have $X_2 = 1$. Analogously, the probability of *not*

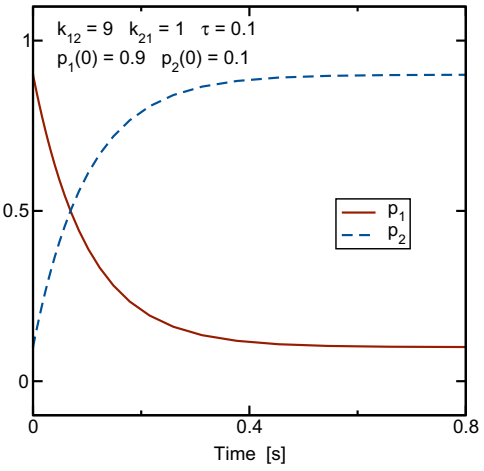

Fig. 4. The evolution of $p_1(t)$ and $p_2(t)$ starting from the (arbitrary) initial condition $p_1(0) = 0.9$ and $p_2(0) = 0.1$. For $t \gg \tau$ one obtains the stationary solution $p_1(\infty) = 0.1$ and $p_2(\infty) = 0.9$, which is reached regardless of the initial condition.

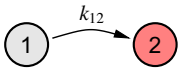

Fig. 5. State transition rate diagram for the calculation of the first passage time from state 1 to state 2, τ_{12}.

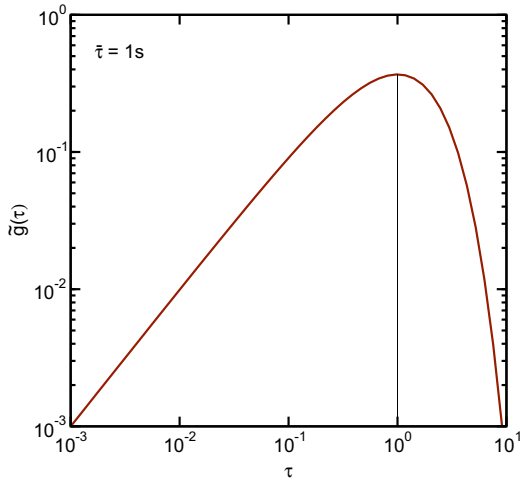

Fig. 6. The exponential distribution (14) with $\bar{\tau} = 1\ s$ on a logarithmic time scale.

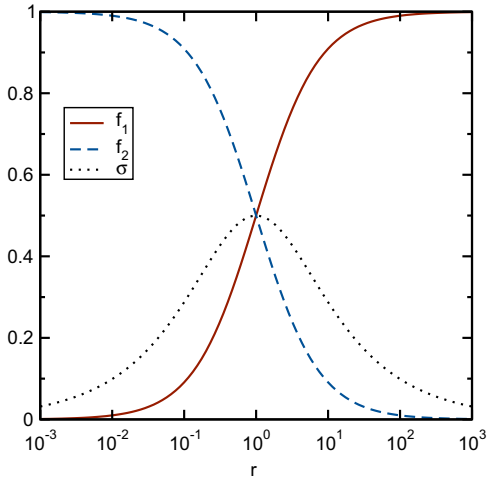

Fig. 7. The expectation values and variance as a function of $r = k_{21}/k_{12}$ under stationary conditions.

being in state 2 is given by $1 - p_2 = p_1$, where we have $X_2 = 0$. Consequently, the expectation value of X_2 is given by

$$\mathrm{E}\{X_2(t)\} = 0 \times p_1(t) + 1 \times p_2(t) = p_2(t). \tag{16}$$

The same result is obtained for all higher-order moments since we only consider occupancies of 0 and 1. So in general we have for all moments

$$\mathrm{E}\{X_i^k(t)\} = \sum_{x=0}^{1} x^k P\{X_i(t) = x\} = p_i(t). \tag{17}$$

In particular, we are interested in the mean behavior, which we shall call f_i and which is given by

$$f_i(t) = \mathrm{E}\{X_i(t)\} = p_i(t). \tag{18}$$

So in this simple example, the mean value equals the probability of being in this particular state. Next, we calculate the variance of the process which is simply

$$\sigma_i^2(t) = \mathrm{E}\{(X_i(t) - f_i(t))^2\} = p_i(t) - p_i^2(t). \tag{19}$$

Under stationary conditions, which are commonly assumed in RTN analysis, the mean and variance are

$$f_2(\infty) = \frac{k_{12}}{k_{12} + k_{21}}, \quad \sigma_2^2(\infty) = \frac{k_{12}k_{21}}{(k_{12} + k_{21})^2}. \tag{20}$$

By introducing the ratio of the transition rates $r = k_{21}/k_{12}$, the mean and variance can be written as

$$f_1 = \frac{r}{1+r}, \quad f_2 = \frac{1}{1+r}, \quad \sigma^2 = \sigma_1^2 = \sigma_2^2 = \frac{r}{(1+r)^2},$$

which is shown in Fig. 7. Note that $\sigma = \sigma_1 = \sigma_2$ holds for all values of r. As can be seen, the standard deviation has a maximum of $\sigma = 0.5$ when $r = 1$, that is, $k_{12} = k_{21}$. Intuitively, provided $\bar{\tau}_c = 1/k_{12}$ falls within the experimental window, the case $r \approx 1$ provides the optimal condition for the analysis of RTN, since there the number of transitions between the states is maximized. Unfortunately, such defects are the most annoying in applications as they produce a maximum of noise. Conversely, consider the cases $r \to 0$ and $r \to \infty$, for both of which we have $\sigma \to 0$. While defects of this kind seem to be favorable from an application point of view because they presumably do not cause any dynamic interference with the circuit, they are more difficult to analyze and understand. However, these defects will go undetected in conventional RTN analysis but cause device reliability issues as will be shown below.

An important issue to be discussed at length in Section 3 is that the rates depend strongly on the applied bias. For example, the forward rate k_{12} depends about exponentially on the oxide field. Thus, under most bias conditions $p_2(\infty)$ will be either zero or one and the variance thus zero. As a consequence, the defect can only be efficiently analyzed in the rather narrow voltage and temperature window where $k_{12} \approx k_{21}$. Since the rates are distributed over a wide range, most defects will go unnoticed in a stationary analysis, as transitions are unlikely to occur.

As an example assume now the simple two-state defect of Fig. 3 with $k_{12} = 1/9\ s^{-1}$ and $k_{21} = 1/1\ s^{-1}$. Since the backward rate k_{21} is larger than the forward rate k_{12}, the defect can be expected to be more likely in state 1. Since from (7) the asymptotic decorrelation time is calculated as $\tau = 0.9$ s, we can expect the occupation probability to be $f_2(\infty) = 1/10$ after a few seconds, irrespective of the state the defect was in initially. Furthermore, the standard deviation will be $\sigma_2(\infty) = 3/10$. A few simulated example realizations of the Markov process together with the estimated mean and

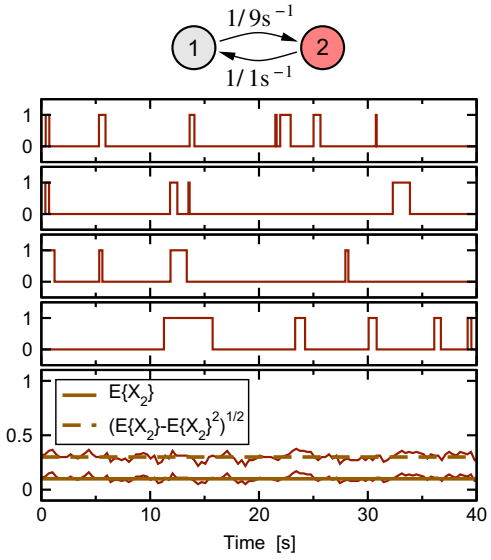

Fig. 8. Stationary behavior of a defect with $k_{12} = 1/9\ s^{-1}$ and $k_{21} = 1/1\ s^{-1}$, which is predominantly in state 1. The top four panels show simulated example realizations of $p_2(t)$, while the thick lines in the bottom panel show the theoretical expectation value, 1/10, and variance, 3/10. The noisy lines are extracted by averaging 100 realizations.

standard deviation are shown in Fig. 8. At this point it is worthwhile to remark that such Markov processes are extremely simple to simulate, requiring a computer program of only a few lines of code, see Ref. [55] for an excellent introduction. Next, lets assume that at a different bias the forward rate decreases from $k_{12} = 1/9\ s^{-1}$ to $k_{12} = 1/999\ s^{-1}$. Under stationary conditions, the defect will then practically always be in state 1, with occasional visits to state 2, see Fig. 9, making it much more difficult to characterize.

2.4. Bias switches

We have seen so far that a defect can be most efficiently characterized when both the capture and emission times fall well into the experimental window and r is close to unity. In reality, this will only be the case for a small number of defects when the transistor is operated at a certain bias and temperature. By scanning a wider bias and/or temperature range, a different set of defects will be active owing to the strong bias-dependence of the time constants. Thus, rather than passively waiting for the defects to spontaneously capture and emit a charge, a state change can be enforced via an external bias switch.

Consider switches of the gate voltage from a low-level V_G^L to a high-level V_G^H, where we assume $|V_G^L| < |V_G^H|$. As the capture time depends about exponentially on V_G, there will be a certain number of defects with $r(V_G^L) \ll 1$ and $r(V_G^H) \gg 1$. These defects are effectively uncharged at V_G^L and become charged at V_G^H. At both V_G^L and V_G^H one has $\sigma \ll 1/2$, that is, these defects will not produce detectable RTN. However, this class of defects is responsible for BTI. In particular, following the switch to the high-level, the defects become charged and at one point during the charging process σ will take on its maximum value, $\sigma = 1/2$. Conversely, following the switch to the low-level, the defects become discharged and σ will again take on its maximum value during this transient.

As an example, assume that at time $t = 0$ s a defect with $k_{12} = 1/9\ s^{-1}$ and $k_{21} = 1/1\ s^{-1}$ has been unperturbed for a time much longer than τ, meaning that the Markov process is stationary. Under stationary conditions the expectation value will be independent of time and, as shown before in Fig. 8, $f_2(0) = 1/10$, while the standard deviation will be $\sigma_2(0) = 3/10$. Then, at $t = 0$ s, the bias conditions are changed, which is by way of example assumed to make the forward rate 81 times larger, that is, $k_{12} = 9/1\ s^{-1}$. According to (6), we expect an exponential transition from $f_2(0) = 1/10$ to $f_2(\infty) = 9/10$, with a time constant of $\tau = 0.1$ s. Following the transition, the standard deviation will settle back to $\sigma_2(\infty) = 3/10$ at the end of the transitionary phase. This is shown in Fig. 10. Still, while $f_2(t)$ moves from 1/10 to 9/10, the standard deviation will reach its maximum of 1/2 when $f_2(t) = 1/2$.

The above result is generally valid: even for defects with $f_2(0) \approx 0$ and $f_2(\infty) \approx 1$ and thus zero variance before and after the bias change, the standard deviation will be maximal at some point during the transition as shown in Fig. 11. These enforced transitions thus provide a convenient opportunity for the characterization of a much larger class of defects, see Fig. 12. It also forms the backbone of the time-dependent defect spectroscopy (TDDS) discussed in Section 3.2.

The switching experiment can be generalized to arbitrary switching sequences, see Fig. 13. For example, the TDDS employs a switch from V_G^L to V_G^H at $t = t_0$ and back to V_G^L at $t = t_1$. Prior to the switch we assume the Markov process to be stationary with a certain probability of being in state 2, given by $p_2(t) = p_2^L$. Then, during the stress or charging period ($t_0 < t < t_1$) we have from (6)

$$p_2(t) = p_2^H + (p_2^L - p_2^H)e^{-t_s/\tau_c}, \tag{21}$$

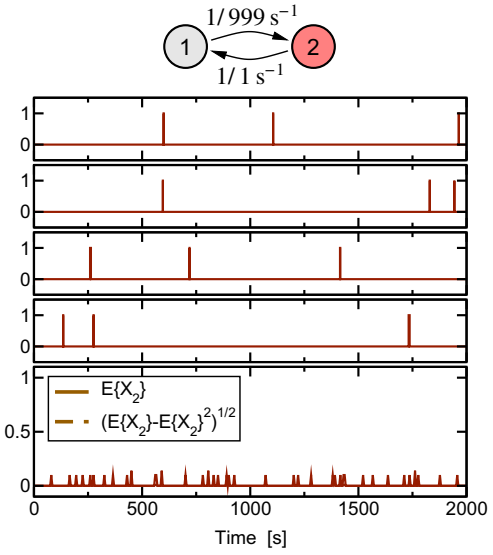

Fig. 9. Same as Fig. 8, but now with $k_{12} = 1/999\ s^{-1}$ and $k_{21} = 1/1\ s^{-1}$. Except for occasional 'bursts', the defect is nearly always in state 1. Such a defect is much harder to characterize than the one shown in Fig. 8.

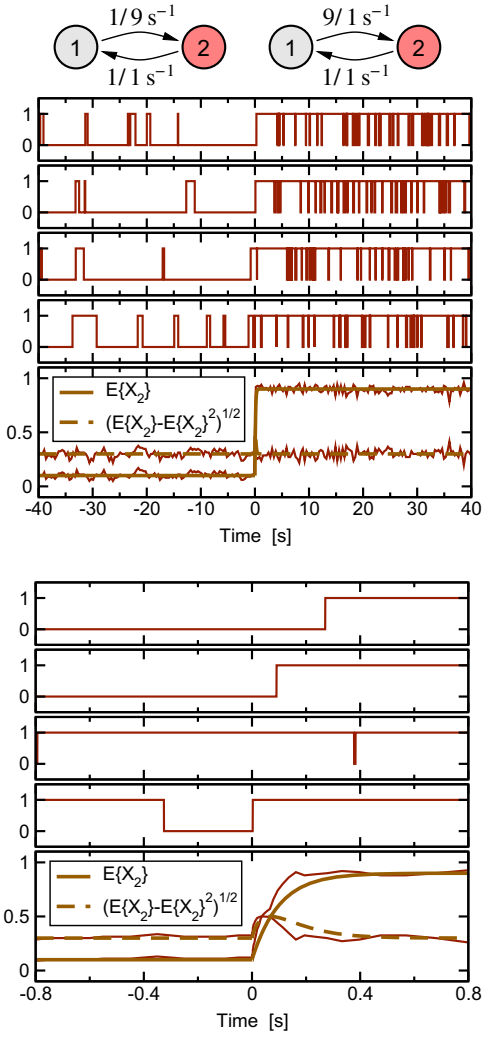

Fig. 10. At time $t = 0$ s, the transition rates of the defect are assumed to change rapidly from $1/9\ s^{-1}$ to $9/1\ s^{-1}$. A zoomed-in version is shown at the bottom. Following the switch, σ briefly goes to 1/2 before returning to its stationary value of 3/10.

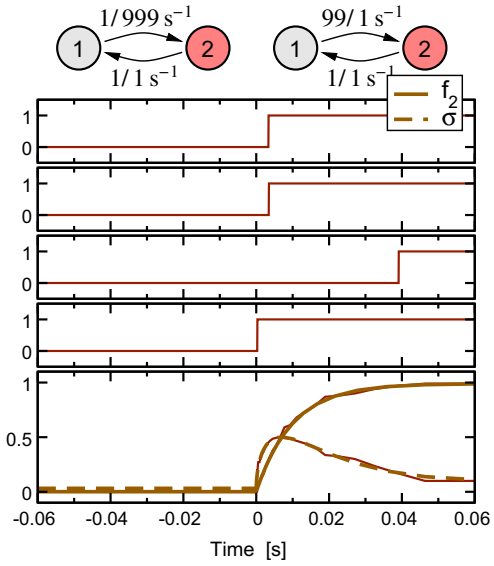

Fig. 11. Same as Fig. 10, but with much smaller k_{12} prior to the switch and larger k_{12} following the switch. Both prior to the switch ($t = 0$) and at $t \to \infty$ we have $\sigma \approx 0$, while $\sigma = 1/2$ during the transition point when $f_1(t) = f_2(t)$. This corresponds to the optimum detection point.

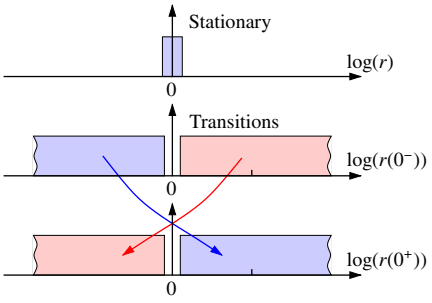

Fig. 12. Sketch of the defect range accessible by experiment. In a stationary setting, as employed in RTN measurements, defects with $r \approx 1$ are most easily accessible (top panel). By changing the bias conditions at $t = 0$, many defects with $r(0^-) < 1$ and $r(0^+) > 1$ (or vice versa) become accessible. Defects with $r(0^-) < 1$ are dominantly seen in TDDS experiments.

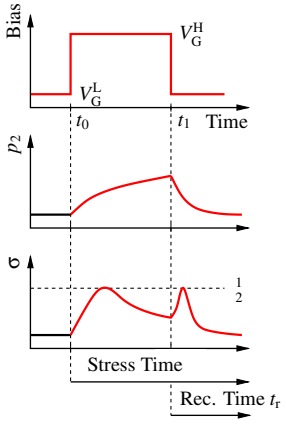

Fig. 13. Defect occupancy and standard deviation during a switch from V_G^L to V_G^H and back.

where we introduced the stationary value of p_2 at V_G^H, $p_2^H = p_2^H(\infty)$, which will be approached as time progresses. The relative time scale is given by the stress time $t_s = t - t_0$.

When V_G is switched back to V_G^L at $t = t_1$, we enter the relaxation or discharging period and find

$$p_2(t) = p_2^L + (P_c - p_2^L)\, e^{-t_r/\tau_e} \tag{22}$$

with $P_c = p_2(t_1)$, the probability of the defect being in state 2 at $t = t_1$, and the discharge or relaxation time $t_r = t - t_1$. Note that the time constant τ is different at the low and high bias. For instance, at high-bias we have with (7)

$$\tau_c = 1/\left(k_{12}^H + k_{21}^H\right) \approx 1/k_{12}^H = \bar{\tau}_c\left(V_G^H\right) \tag{23}$$

which is usually dominated by the average capture time when emission is negligible. Conversely, at low-bias we have

$$\tau_e = 1/\left(k_{12}^L + k_{21}^L\right) \approx 1/k_{21}^L = \bar{\tau}_e\left(V_G^L\right), \tag{24}$$

when capture is negligible. In general, however, particularly when the capture and emission times are scanned over a wide bias range, the full expression (23) and (24) have to be used [56].

2.5. Impact on the threshold voltage

So far we have summarized methods which can be used to model the random transitions occurring between the various defect states. The next question to answer is how the charge state of the defect impacts the device behavior. It is normally assumed that neutral defects do not interact with the remainder of the device, while charged defects cause a shift in the threshold voltage and degrade the mobility. In general, the impact on threshold voltage and mobility is detrimental, that is, reduces the drain current delivered by the transistor. Only occasionally an improvement is observed [57]. In the following we will restrict our discussion to the impact of charged defects on the threshold voltage.

Usually, one differentiates between fast interface states and slower border traps, both of which contribute to an effective area density of charge located at the Si–SiO_2 interface. Since the defects randomly change their occupancy with time, the time-dependent effective threshold voltage can be written as

$$V_{th}(V_G, t) = V_{th0} - \frac{Q_{it}(V_G, t) + Q_{ox}(V_G, t)}{C_{ox}}, \tag{25}$$

with $C_{ox} = \epsilon_0 \epsilon_r / t_{ox}$, the capacitance per area and t_{ox} the oxide thickness. Since interface states have time constants in the sub-microsecond regime, the transitions between their states cannot be experimentally resolved. Rather, experiments record their average occupancy, that is the expectation value of being in the charged state given by f_2. When the bias is changed, they respond rapidly to changes in the Fermi-level, without detectable transients, so that $Q_{it}(V_G, t)$ can be written as $Q_{it}(V_G)$. For instance, it is the change of $Q_{it}(V_G)$ with V_G which causes a change of the subthreshold slope compared to the ideal MOS transistor with $Q_{it} = 0$.

Most border traps, on the other hand, are too slow to quickly follow the bias change and will produce notable transients according to (6). As such, for a single donor-like trap we can write

$$Q_{ox}(V_G, t) = q \frac{1 - x/t_{ox}}{WL}\, \eta_r\, f_2(t), \tag{26}$$

where $f_2(t) = p_2(t)$ according to (21) and (22). The depth of the defect into the oxide is x, the oxide thickness is t_{ox}, while W and L are the channel width and length of the transistor. The above equation is derived from Gauss' law under the simplifying assumption that the charge of the defect is homogeneously spread across the oxide, thereby forming a charge sheet [58]. This assumption is not correct, since the real three-dimensional distortion of the potential caused by the defect charge can lead to a considerably larger impact on V_{th} [59]. In order to consider the deviation from the charge sheet

approximation, we introduce the empirical parameter η_r, which can have values up to 10 [59–62]. Thus, the impact of a single defect on V_{th} is given by

$$\frac{Q_{ox}(V_G,t)}{C_{ox}} = \eta f_2(t), \tag{27}$$

with $\eta = \eta_0\eta_r$ and the step height of the charge sheet approximation

$$\eta_0 = q\frac{1 - x/t_{ox}}{WLC_{ox}}. \tag{28}$$

Prior to a bias switch, say at V_G^L, the defect occupancy is stationary and will produce RTN. This RTN can sometimes be analyzed, but most of the time it will be outside the experimental window. Particularly when monitoring a large number of defects, the analysis of the complicated RTN is not possible. As such, we cannot easily record f_2^L, since it is already implicitly contained in the 'unshifted/reference' V_{th}. The shift in V_{th} is thus given as

$$-\Delta V_{th}(t) = \eta(f_2(t) - f_2^L) = \eta\Delta f_2(t).$$

Consequently, as the above is our reference value, we have $\Delta f_2(t) = 0$ for $t < t_0$. During stress ($t > t_0$) we have with (21)

$$-\Delta V_{th}(t_s) = \eta a(1 - e^{-t_s/\tau_c}), \tag{29}$$

with $a = f_2^H - f_2^L$. During recovery ($t > t_1$) the defect will discharge like (22) as

$$-\Delta V_{th}(t_s,t_r) = \eta a(1 - e^{-t_s/\tau_c})e^{-t_r/\tau_e}. \tag{30}$$

Note that ΔV_{th} is a stochastic quantity with the above giving its expectation value. When recorded experimentally, ΔV_{th} can be anything between ηa and zero, but when averaged over a large number of experiments, (30) will be obtained, see Fig. 13.

2.6. Substitute circuit

In order to analyze the temporal behavior of the defect occupation probabilities, substitute circuits which formally result in the same differential equation can be used [32]. For illustration purposes we restrict ourselves to uncharged defects at $t = 0$ s which fully charge as time progresses. Then, the temporal evolution of the expectation value and variance are

$$f_2(t) = 1 - e^{-t/\tau} \quad \text{and} \quad \sigma^2(t) = e^{-t/\tau} - e^{-2t/\tau}.$$

The maximum of the standard deviation is observed at about τ, at $t_{max} = \tau\ln(2)$, where

$$f_1(t_{max}) = f_2(t_{max}) = \sigma(t_{max}) = 1/2.$$

In other words, at that particular point during the transition where f_1 equals f_2, the standard deviation always peaks at 1/2, see Fig. 14.

2.7. Multi-state defects

Experiments show that defects can have more than two states. An example is given in Fig. 15, where a defect produces RTN following NBTI stress. This RTN is only temporary and stops after a while, which is why it has been termed temporary RTN [3,4]. Such temporary RTN can also occur spontaneously, without prior application of a stress voltage, a phenomenon known as anomalous RTN [52]. The logical explanation for both phenomena is the existence of metastable states, which will be discussed now.

The previous discussion on two-state defects can be easily generalized to a defect having N states. Let $X_i(t)$ be the occupancies of each state, with $X_i(t) = 1$ when the defect is in state i and $X_i(t) = 0$ otherwise. Since the defect can only be in one of its states, it follows from $\sum_i X_i(t) = 1$ that $X_j(t) = 0\ \forall j \neq i$ when $X_i(t) = 1$.

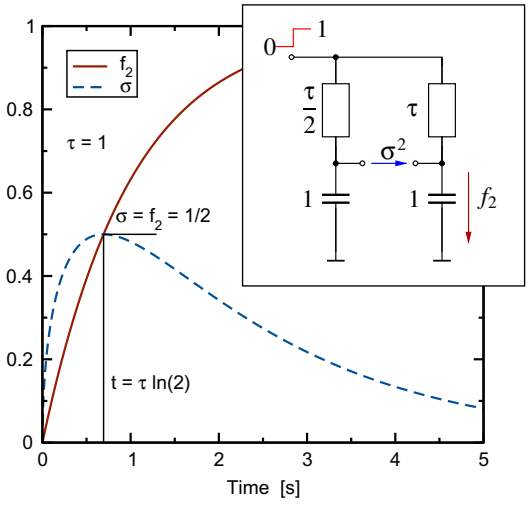

Fig. 14. Substitute circuit for the calculation of f and σ.

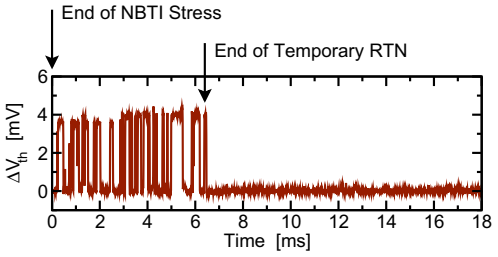

Fig. 15. Example of a multi-state defect: following NBTI stress, the defect produces RTN only for a limited amount of time.

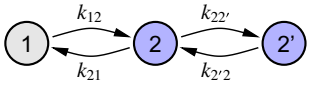

Fig. 16. Defect with three states to explain anomalous RTN. State 1 is assumed to be electrically neutral, while both state 2 and state 2′ are assumed to carry a single negative charge. As such, states 2 and 2′ are electrically indistinguishable.

The conditional probabilities describing a transition within the next infinitesimally time interval h are given by

$$P\{X_j(t+h) = 1 | X_i(t) = 1\} = k_{ij}h + O(h),$$
$$P\{X_i(t+h) = 1 | X_i(t) = 1\} = 1 - \sum_{j\neq i} k_{ij}h + O(h).$$

Just as in the simple case of two defect states, the Master equation describing the transitions is obtained by considering the limit $h \to 0$ as [63,53]

$$\frac{\partial p_i(t)}{\partial t} = -p_i(t)\sum_{j\neq i} k_{ij} + \sum_{j\neq i} k_{ji}p_j(t). \tag{31}$$

Note that due to $\sum_i p_i(t) = 1$, only $N - 1$ equations are linearly independent.

Although the analytic solution of the Master equation is in principle straight-forward to obtain, it is somewhat lengthy and unintuitive. The primary reason for that is that contrary to the simple solution of the two-state defect, the mere addition of one single state dramatically increases the complexity of the system behavior. Also, the behavior of the system changes considerably depending on the initial conditions and the choice of the transition rates. However, since the relevant parameter range does not require us

to explore the most general solution, we will restrict our discussion to the most important cases observed experimentally.

We start with the multi-state defect as observed by Uren et al. [52]: while studying RTN in nMOSFETs, they observed a defect which produced regular RTN only for stochastic amounts of time. These RTN periods were followed by periods where the defect remained negatively charged. This behavior was termed *anomalous RTN* and explained by the defect having one additional metastable state. In order to model such a behavior, we consider a defect with one additional metastable state attached to state 2, which we shall denote by 2′. In state 1 the defect is neutral, while both 2 and 2′ are negative. Since observable RTN requires a change in the charge state of the defect, the rates between state 1 and 2 must be larger than the rates between the state 2 and 2′, see Fig. 17.

Similarly to the anomalous RTN defect in nMOS transistors, a phenomenon termed *temporary RTN* has been recently observed in pMOS transistors [3,4]: following a charging pulse, a defect was activated which produced RTN for a limited amount of time before the signal vanished. Such a phenomenon can be described by a similar defect model, but this time state 2 is positively charged while 1′ and 1 are neutral. An example case is shown in Fig. 18, where the defect is assumed to be in state 2 at $t = 0$ s. For the first few seconds the defect switches back and forth between states and 2 and 1′. Once the transition to state 1 is made, the defect remains inactive for about $1/k_{11'} \approx 10^6$ s with the parameters given in Fig. 15. Note that application of a stress bias increases $k_{11'}$ by many orders of magnitude, thereby allowing for a transition to state 2 again.

2.8. First passage time for a three-state defect

While already a defect with three states can show rather complicated behavior, a particular case is of fundamental importance: most defects appear to be neutral at low bias conditions and become charged at higher bias. This implies that at low bias the neutral state is the equilibrium state while at high bias the rates change in such a way that the charged state becomes the new equilibrium state, see Fig. 19. In such a case the description of the transitions is considerably simplified and can often be approximated by an 'effective' two-state defect, as will be outlined in the following.

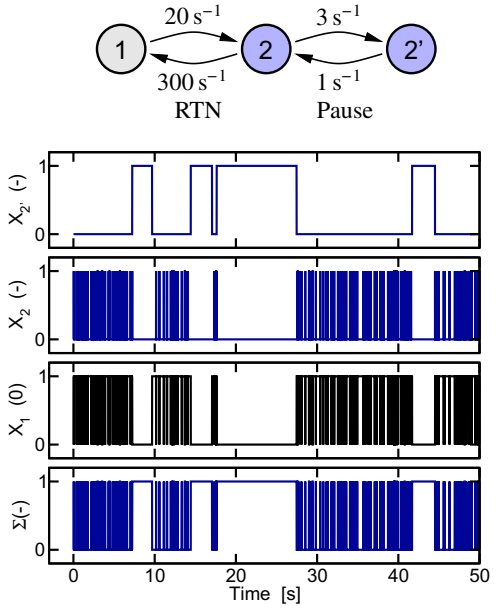

Fig. 17. Anomalous RTN as produced by the defect of Fig. 16. Whenever the defect goes into the metastable state 2′, the RTN stops and the defect remains in the negatively charged state.

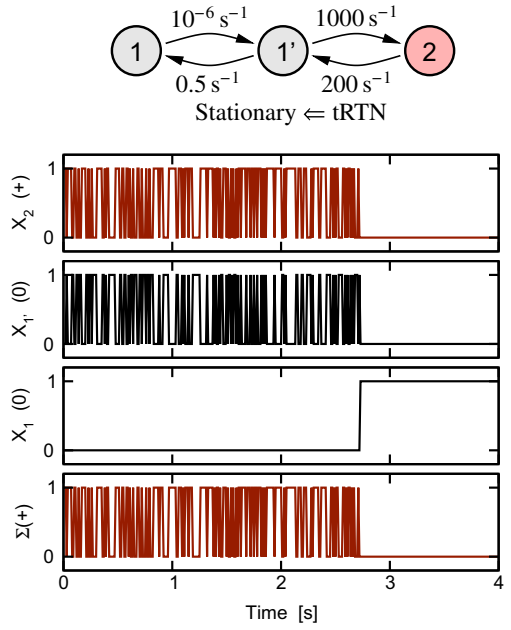

Fig. 18. Temporary RTN starting from the defect being in an excited state 2. Once the defect goes back into the equilibrium state 1, the tRTN stops and the defect remains electrically neutral.

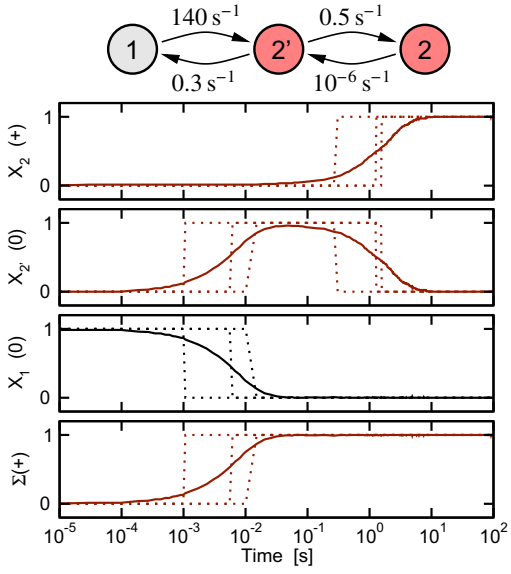

Fig. 19. Transition from the initial state 1 to the equilibrium state 2 via the metastable transition state 2′. The dashed lines give three example traces while the solid line is an average over 200 such traces.

We start by calculating the first passage time for a general three-state defect such as shown in Fig. 16. The states have been labeled A, B, and C and we determine the probability distribution of the first passage from state A to state C. Since the backward rate from state C to state B does not influence the first passage time, it can be omitted in the analysis, see Fig. 20. The Master equation for the problem is

$$\frac{dp_A}{dt} = -bp_A + ap_B,$$
$$\frac{dp_B}{dt} = bp_A - ap_B - cp_B,$$
$$\frac{dp_C}{dt} = cp_B.$$

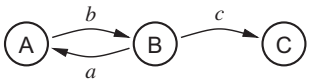

Fig. 20. Simplified Markov chain for the calculation of the first-passage time from state A to state C.

In general, the Master equation of such multi-state defects is a first-order linear ordinary differential equation system, which can be solved using standard methods to find

$$p_C(t) = 1 - \frac{1}{\tau_2 - \tau_1}(\tau_2 e^{-\tau/\tau_2} - \tau_1 e^{-\tau/\tau_1}), \tag{32}$$

with

$$\tau_1^{-1} = \frac{1}{2}\left(s + \sqrt{s^2 - 4bc}\right), \tag{33}$$

$$\tau_2^{-1} = \frac{1}{2}\left(s - \sqrt{s^2 - 4bc}\right). \tag{34}$$

and $s = a + b + c$. Retracing the derivation used to calculate the first passage time for the two-state defect, we see that the probability that the defect is not yet in state C is given by $P\{\tau < t\} = p_C(\tau)$. Thus, the probability density function of the first passage time from state A to state C of the three-state defect is

$$g(\tau) = \frac{dp_C(\tau)}{d\tau} = \frac{e^{-\tau/\tau_2} - e^{-\tau/\tau_1}}{\tau_2 - \tau_1}. \tag{35}$$

This is the normalized difference between two exponential distributions, where the faster distribution with mean τ_1 'truncates' the slower distribution with mean τ_2 below $\tau < \tau_1$. The expectation value of τ is simply

$$\bar{\tau} = E\{\tau\} = \int_0^\infty \tau g(\tau)\, d\tau = \tau_1 + \tau_2 = \frac{a + b + c}{bc}. \tag{36}$$

For $a = 0$ (no back transition for B to A), the two auxiliary time constants reduce to $\tau_1 = 1/b$ and $\tau_2 = 1/c$. In such a case the first passage time is simply given by the two sequential steps from A to B and from B to C. Also, one can easily convince oneself that for large b (rapid transition from A to B) and $a = 0$, (35) reduces to a simple exponential distribution.

In general, however, we have $\tau_1 \geqslant 1/b$ and $\tau_2 \geqslant 1/c$, that is, both effective transition times are made larger by the back transition from B to A.

Following (14), the p.d.f. transformed onto a logarithmic scale is $\tilde{g}(\tau) = \tau g(\tau)$, which is shown in Fig. 21. As can be seen, for smaller values of τ_1, the p.d.f. of the first passage time of the three-state defect is very close to an exponential distribution with parameter $\bar{\tau} = \tau_1 + \tau_2$. While the expectation value of the exponential distribution is per construction exactly the same as that of the three-state p.d.f., this is only approximately true for the higher-order moments. This is because the first transition from A to B, although assumed to be fast, imposes a lower limit on the fastest overall transitions by reducing the density in the left tail of the probability density. Still, for sufficiently small τ_1, it appears than an effective two-state model can be constructed which represents this particular transition of the three-state defect reasonably well.

We now proceed to construct an effective two-state model of the three-state defect for the important special case of a switch from V_G^L to V_G^H and back, see Fig. 22. At V_G^H, we are primarily concerned with the transition from state 1 to state 2, where the latter is assumed to be the equilibrium state at this bias. Under these conditions, the first passage time is the average capture time which is obtained from (36) as

$$\bar{\tau}_c = \frac{k_{2'1} + k_{12'} + k_{2'2}}{k_{12'}\, k_{2'2}}. \tag{37}$$

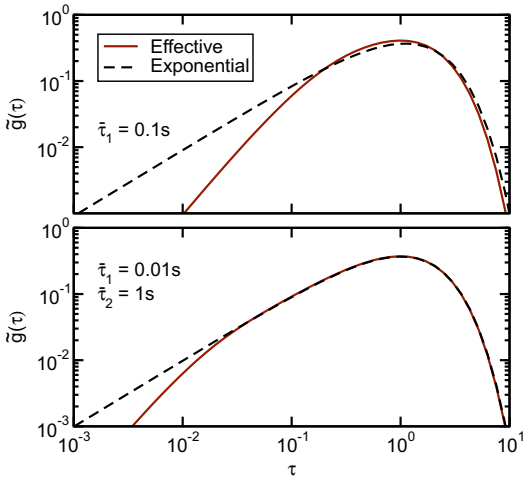

Fig. 21. Impact of τ_1 on the shape of the distribution of the first passage time from state A to C. With decreasing τ_1, the distribution becomes increasingly closer to exponential distribution.

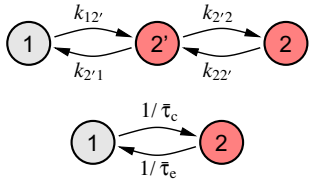

Fig. 22. Top: Defect with three states consistent with the TDDS experiments on pMOS transistors. State 1 is assumed to be electrically neutral, while both state 2 and state 2′ are assumed to carry a single positive charge. As such, while states 2 and 2′ are electrically indistinguishable, state 2′ impacts the capture time. **Bottom**: Effective two state model of the three-state defect obtained by stochastic model order reduction.

Analogously, at V_G^L, the transition of interest is from state 2 to state 1. Again, from (36) we now obtain the corresponding average emission time as

$$\bar{\tau}_e = \frac{k_{2'2} + k_{22'} + k_{2'1}}{k_{22'}\, k_{2'1}}. \tag{38}$$

Eqs. (37) and (38) will be used in Section 7 to understand the bias- and temperature-dependence of the experimentally observed data.

3. Experimental

The theoretical concepts elaborated in the previous sections will now be compared against experimental data. Discrete charging and discharging events can be observed in small-area transistors, as there only a few defects are active. By carefully selecting the bias conditions and stress time, a single defect can be charged and discharged in a controlled manner. Starting from the experimental observation of the stochastic charging/discharging transients of a single trap, we later proceed to the study of recovery traces showing a handful of defects obtained from subjecting the device to harsher stress conditions.

3.1. Single trap

A simple experimental validation of the stochastic nature of charge capture and emission can been obtained by monitoring discharging transients of a single defect in a pMOS [64–66,61]. Experimentally, it is much easier to monitor the discharging rather than

the charging transients with sufficient accuracy, because at lower $|V_G|$ the impact of a single charge on I_D is much larger [66].

The experiment proceeds as shown previously in Fig. 13: After a charging period of duration t_s at a higher gate voltage, V_G^H, the gate voltage is switched to a lower discharging voltage V_G^L during which the subsequent charge emission transient is recorded. The charging/discharging sequence is repeated N times until accurate statistics have been gathered. A few selected recovery traces which clearly show a single discharging event are shown in Fig. 23. From the 100 discharging transients recorded, about 70 did not contain a discrete step, meaning that no charge had been trapped in the charging step. The remaining 30 traces showed an initial threshold voltage shift of $\eta \approx 1.5$ mV which disappeared after a few seconds. Since no RTN is observed at V_G^L, we conclude that the initial defect occupancy prior to the switch is $f^L \approx 0$. Also, at V_G^H no RTN is observed, thus $f^H \approx 1$. From this, according to (30), the maximum observable change in the *average* occupancy is $a = 1$. Furthermore, for a defect which approximately follows first-order kinetics, the averaged discharge transient must be exponential, where for $t_r = 0$ we have

$$P_c = -\frac{\Delta V_{th}(t_s, 0)}{\eta} = 1 - e^{-t_s/\tau_c}, \tag{39}$$

where P_c is the probability that the defect is charged after t_s at V_G^H. From the experimental observation $P_c = 0.3$ we obtain the average capture time by inverting (42) as $\tau_c \approx 3$ ms. The average emission time, on the other hand, is obtained by matching the averaged relaxation traces to

$$P_e = -\frac{\Delta V_{th}(t_r)}{\eta P_c} = e^{-t_r/\tau_e}, \tag{40}$$

which gives $\tau_e \approx 4$ s. It is worth recalling that the discrete step height η is not directly visible in the averaged curve of Fig. 23, which shows ηP_c.

It has to be kept in mind that the extracted values of τ_c and τ_e do not necessarily correspond to τ_c^H and τ_e^L, but contain in general also contributions from τ_e^H and τ_c^L, see (23) and (24). However, since $f^L \approx 0$ and $f^H \approx 1$, we conclude from Eqs. (23) and (24) that in this case $\tau_c^H \approx \tau_c$ and $\tau_e^L \approx \tau_e$ holds. For experimental data on the more general case, see [67].

As will become apparent when discussing the theoretical properties of charge capture and emission, it is interesting to note here that this particular defect has an average capture time of 1 ms while the emission time is 4 s, a factor 4000 larger.

3.2. Multiple traps

The procedure outlined above can be applied to the same defect at various stress and recovery voltages as well as temperatures in order to extract the bias and temperature dependence of the capture and emission times. While such an approach is feasible and already provides a wealth of information [68,66], it requires careful selection of devices with only a single active defect in the whole bias and temperature range. This is a considerable restriction reminiscent of RTN analysis, although not nearly that restrictive as the bias switches considerably increase the number of visible defects. Still, at larger stress times or stress voltages, inevitably multiple defects become charged, hampering the manual analysis. In order to efficiently analyze a larger number of defects, the *time-dependent defect spectroscopy* (TDDS) has been suggested [3,4].

The basic TDDS setup is the same as discussed above for the study of individual defects. The primary difference lies in the analysis of the recovery traces which are no longer averaged but analyzed individually. In the initial analysis step, the measured recovery traces are approximated by discrete steps according to

$$-\Delta V_{th}(t_s, t_r) = \sum_k \eta_k H(t_s - \tau_{c,k}) H(\tau_{e,k} - t_r), \tag{41}$$

using a straight-forward curve tracing algorithm, with $H(x)$ being the unit step function. Emission events are characterized by their emission time and step height, $(\tau_{e,k}, \eta_k)$, which are then binned into a 2D histogram, see Fig. 24. The entries in the 2D histogram are then normalized by N to obtain the spectral map after the stress time t_s. In (41) the re-capture of charge during one transient is ignored, which is normally the case when $\bar{\tau}_c\left(V_G^L\right) \gg \bar{\tau}_e(V_G^L)$. The latter can be experimentally assured by selecting devices which do not show appreciable RTN at $V_G = V_G^L$.

The spectral maps do not only allow for the extraction of $\bar{\tau}_e$ and the step-height η, but also give detailed information on the capture time constant $\bar{\tau}_c$. Naturally, determination of $\bar{\tau}_c$ is only possible for

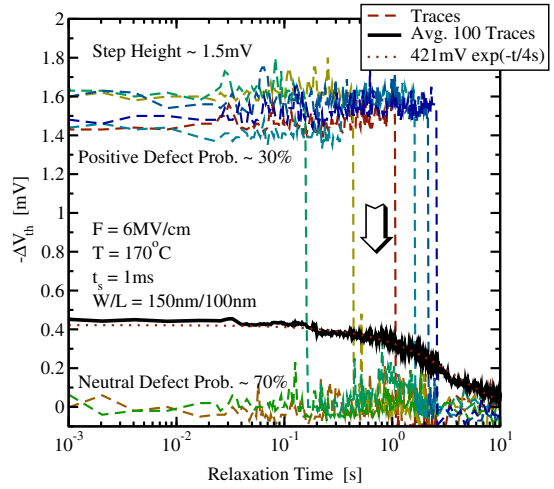

Fig. 23. Experimental recovery traces of a defect with a characteristic step-height of 1.5 mV. The probability that the defect is positively charged after a stress at 5 MV/cm for 1 ms is about 30%, implying that 70% of the traces show no signal in this window (only 2 examples shown). Averaging results in the expected $\exp(-t/\tau)$ behavior. Note that a 1 ms stress introduces a positive charge which recovers only after 4 s.

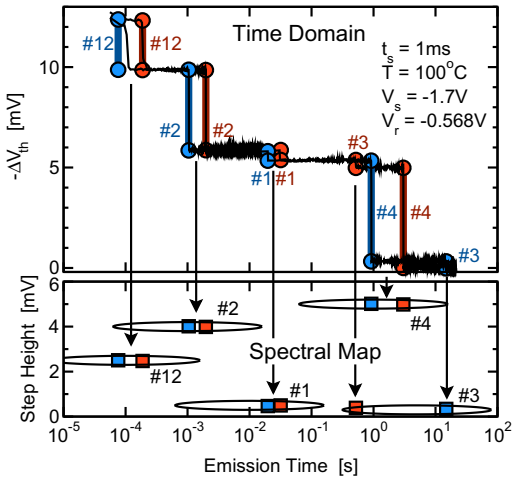

Fig. 24. The TDDS: Two typical ΔV_{th} recovery traces of a small-area pMOSFET. The measured data are given by the noisy black lines (top). The thick blue and red lines together with the symbols mark the emission times and step heights, unambiguous fingerprints of each defect which constitute the spectral map (bottom). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

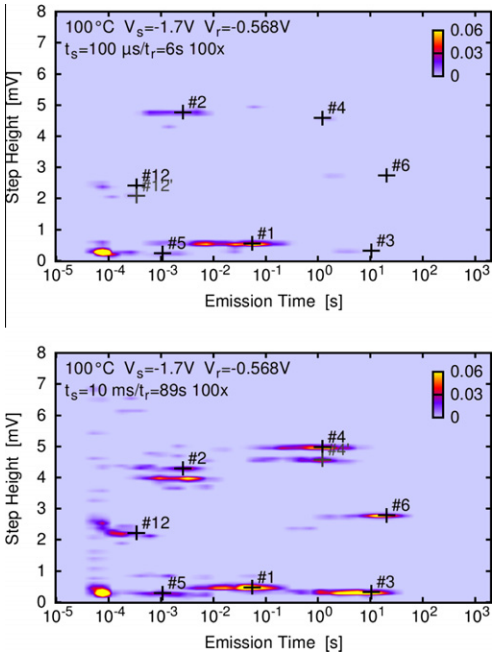

Fig. 25. Two TDDS spectral maps at two stress times, $t_s = 100\,\mu s$ (left) and $t_s = 10$ ms (right). With increasing stress time, the number of defects in the map increases. The width of each cluster is given by the exponential distribution of τ_e (considered on a log-scale) and the extracted defects/clusters are marked by '+'.

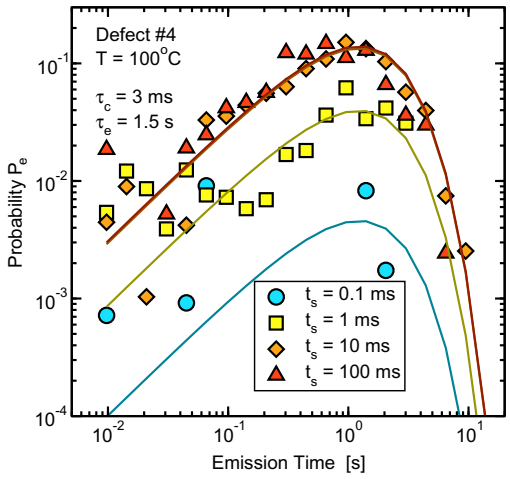

Fig. 26. Extraction of the emission time constant for defect #4 using a 1D histogram where all events with $4\,mV < \eta < 6\,mV$ are plotted as a function of the emission time. The data is shown for four different stress times. At each stress time, the data (symbols) can be described by $P_c P_e$ (lines). For $t_s \geqslant 10$ ms the P_c equals 1, that is, the number of emission events does not change anymore with increasing stress time.

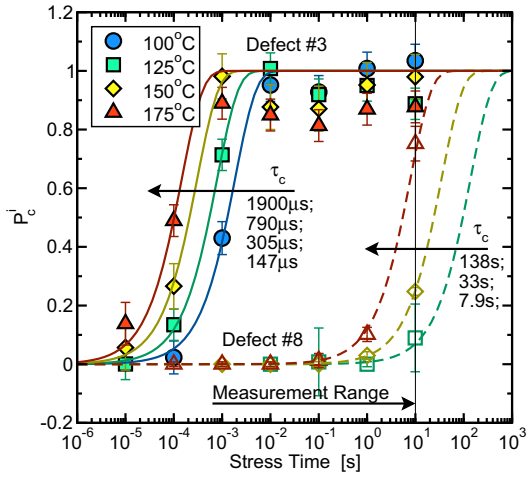

Fig. 27. Example extraction of the capture time constant at four different temperatures. With increasing stress time and temperature the number of defects contributing to the map increases. This makes the identification of the discrete steps more difficult and the noise level in the maps increases. Consequently, the clusters become wider, resulting in a spurious decrease of $P_{c,k}$, which may show a visible deviation from unity even for $t_s \gg \bar{\tau}_c$.

defects visible on the spectral map. However, defects with $\bar{\tau}_c$ much smaller than the stress time can be expected to appear with a 100% probability on the spectral map. In such a case no information on $\bar{\tau}_c$ is obtained. On the other hand, defects with capture times much larger than the stress time will not appear at all. In the intermediate regime, however, where the capture time is on the order of the stress time, the number of hits per cluster will increase exponentially, allowing for an accurate determination of $\bar{\tau}_c$. Provided $\bar{\tau}_{c,k}(V_{stress}) \ll \bar{\tau}_{e,k}(V_{stress})$ we have

$$P_{c,k} = 1 - e^{-t_s/\bar{\tau}_{c,k}}. \tag{42}$$

Thus, M spectral maps with increasing stress time are recorded to also extract the capture time constants of the defects appearing on the maps. One map per decade in time proved sufficiently accurate and we used stress times $t_s \in \{1\,\mu s, 10\,\mu s, \ldots, 10\,s\}$. The overall effect of using M maps is that the defects are not only deconvoluted according to their emission times and step-heights as on a single map, but also according to their capture times. Thus, provided that a defect has either a different emission time, a different step-height, or a different capture time compared to any other defect, it can be clearly identified. This is the reason for the unique accuracy of the TDDS.

For demonstration purposes, the result of a supposedly simpler analysis based on 1D histograms is shown in Fig. 26. This particular 1D histogram is an integrated version of the 2D spectral map over the step height in the window $4\,mV < \eta < 6\,mV$. The probability of having an emission event with $\tau_e \in [\tau_j, \tau_{j+1}]$ is given by $P_c P_e$, which nicely agrees with the experimental data as demonstrated in Fig. 26, see also Fig. 6.

In Fig. 27 the extraction of $\bar{\tau}_c$ for defects #3 and #8 is shown based on the P_c extracted from the spectral maps. Defect #3 has about the same step height as #1 (≈0.4 mV) while their emission time constants are separated by a factor of about 500 at 100 °C, which decreases to 10 at 175 °C. At 100 °C, this separation is reasonably large due to the different activation energies (see Fig. 25) and $P_{c,3}(t_s)$ approaches unity, as expected from $P_c(t_s \gg \bar{\tau}_c)$. Since the activation energy of $\bar{\tau}_e$ of #3 is about twice as large as that of #1, the emission events related to #1 and #3 increasingly overlap with increasing temperature, making the extraction of both $P_{c,1}(t_s)$ and $P_{c,3}(t_s)$ more difficult. This is visible as a marked deviation of $P_{c,3}(t_s)$ from unity in Fig. 27. Another interesting case given in Fig. 27 is #8, which has a capture time constant larger than the largest stress time used in our experiments.

As an example, the extracted capture and emission time constants are shown in Figs. 28 and 29, which are clearly temperature activated with an activation energy of about 0.6 eV and depend in a non-exponential manner on the stress bias.

As was already identified decades ago in the context of RTN [49], both charge capture and emission are thermally activated, ruling out an elastic tunneling process but being compatible with nonradiative multiphonon theory [69,47,70]. The temperature dependencies of $\bar{\tau}_e$ and $\bar{\tau}_c$ are shown in Fig. 30, demonstrating a wide spread in the activation energies. We will get back to this later in Section 4.

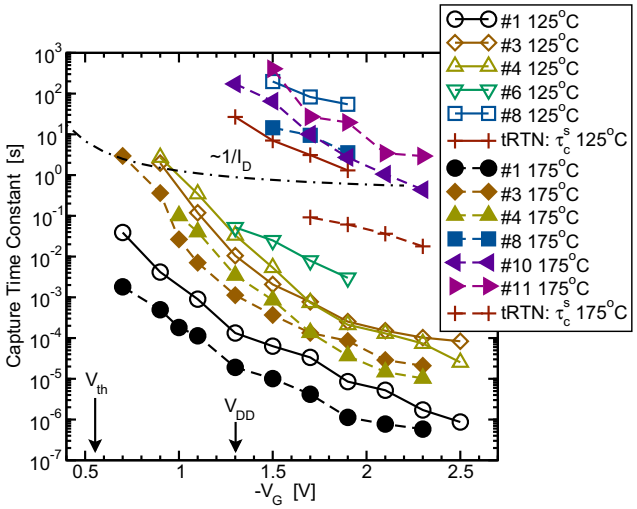

Fig. 28. Voltage and temperature dependence of the capture time constant for seven defects. Defect #6 was visible during the initial experiments only (taken at 125 °C) and then disappeared permanently, #10 and #11 were outside our experimental window at 125 °C. A strong field/voltage dependence clearly different from exponential is observed for all defects. Also shown is the $1/I_D$ tendency expected from the standard SRH model [50], see Section 5, which is considerably weaker.

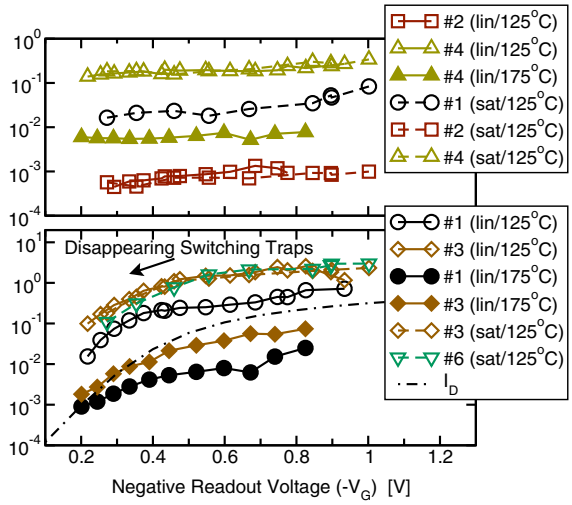

Fig. 29. Gate voltage dependence of the emission time constants at 125 °C and 175 °C, measured in the linear and the saturation regime. Defects #1/sat, #2, and #4 (top) may be also described by a standard model as they only show a weak (if at all) field dependence at lower voltages. In contrast to that, $\bar{\tau}_e$ 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 two-state defect model but indicates the existence of transitionary metastable state, see Section 7.

Finally, the discrete capture/emission time (CET) map constructed from the extracted data is shown in Fig. 31. With increasing temperature, both the capture and emission times become shorter. With increasing stress bias V_G^H, the capture times become shorter. The emission times, which are always recorded at the same V_G^L remain unaffected. While the discrete CET map extracted via the TDDS provides a wealth of information, it is very time consuming to extract. As such, it does not contain enough data to provide statistically relevant information, for instance on the correlation between the capture time constants and their activation energies. It therefore appears more promising to extract that information using CET maps measured on large-area devices [66].

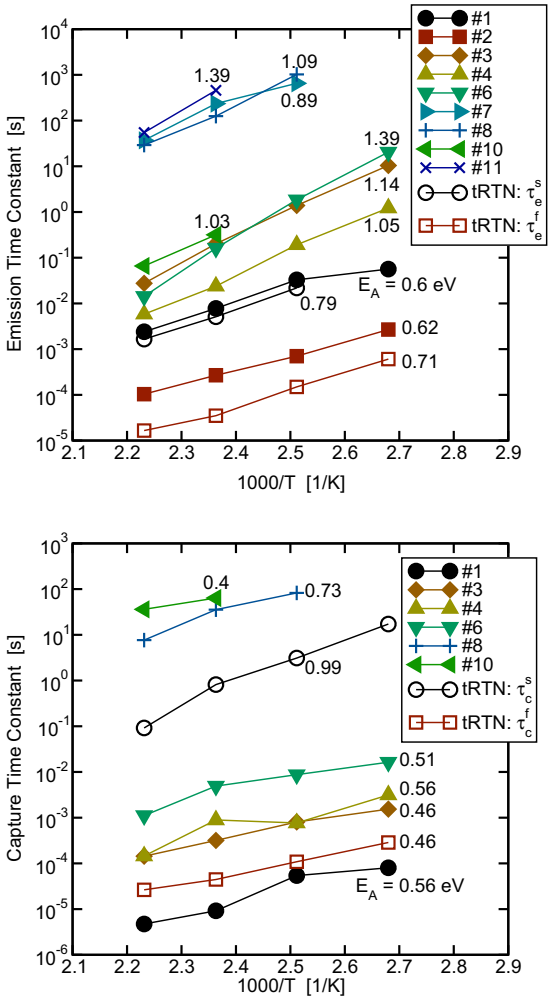

Fig. 30. Top: Arrhenius plot of the emission time constants $\bar{\tau}_e$ for a recovery bias of −0.55 V. The approximate activation energies are given for each defect where a wide spread is observed. **Bottom**: Arrhenius plot of the capture time constant $\bar{\tau}_c$ for a stress bias of −1.7 V.

4. Large numbers of defects

So far we have dominantly considered the response of individual or a handful of defects to switches between two bias levels. In the following, the response of a large number of defects is discussed under the assumption that the defects do not interact. Although this is generally a good approximation, because the defects are on average too widely separated for their wavefunctions to overlap, it has been observed that the occupancy of one defect impacts the step height of another [3] via electrostatic changes in the channel. Furthermore, it is conceivable that these electrostatic changes modify the capture and emission times of surrounding defects [4]. Nonetheless, these interactions will be neglected in the following, as they are neither well documented nor properly understood at the present. Furthermore, while the discussion is limited to donor-like defects as relevant for NBTI, the generalization of the model to acceptor-like defects required for PBTI should be obvious.

Given the previous results, the total shift in the threshold voltage, $\Delta V_{th}(t_s, t_r)$, is due to the collection of N independent forward and backward transitions between a neutral and a charged state. We again consider the simple scenario where the gate voltage is switched between two levels, high and low, with high corresponding to stress and low to recovery. The behavior of each defect is

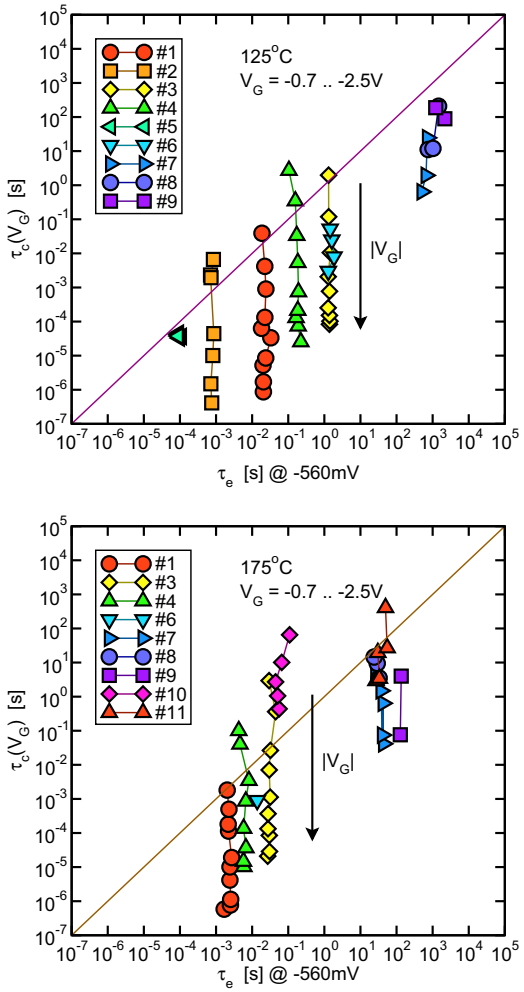

Fig. 31. The discrete capture/emission time map constructed from TDDS data for two temperatures, 125 °C (top) and 175 °C (bottom). With increasing stress voltage V_G^H, the capture times decrease, while the emission time, which is always recorded at the same V_G^L, is not affected.

characterized by its capture and emission time constants, τ_c and τ_e, both of which strongly depend on bias and temperature. Since we only consider switches between two bias levels, we also just need to worry about their values at these two levels which will be denoted as τ_c^H, τ_c^L, τ_e^H, and τ_e^L. For a switch from the low to the high level, the time constant determining the transition follows from (7) as

$$\tau_c = \frac{1}{1/\tau_c^H + 1/\tau_e^H}. \tag{43}$$

Conversely, when the voltage is switched from the high to the low level, the time constant determining the emission process is given analogously by

$$\tau_e = \frac{1}{1/\tau_c^L + 1/\tau_e^L}. \tag{44}$$

Prior to the application of the stress bias we assume that the defects are in the stationary regime. In that case their occupancies are calculated from (7) as

$$f^L = \frac{\tau_e^L}{\tau_e^L + \tau_c^L}. \tag{45}$$

After the gate voltage has been at the high level for an infinitely long time, the defect occupancy is given analogously by

$$f^H = \frac{\tau_e^H}{\tau_e^H + \tau_c^H}. \tag{46}$$

After (21), the transition from f^L to f^H follows

$$f(t_s) = f^H + (f^L - f^H)\ e^{-t_s/\tau_c}.$$

If after a certain stress time t_s the bias is switched back to the low level, the occupancy will go from the value given by $f(t_s) \leqslant f^H$ back to f^L following (22) as

$$f(t_s, t_r) = f^L + (f(t_s) - f^L)\ e^{-t_r/\tau_e}.$$

A stress/relaxation experiment based on the above switching scheme is the simplest case but already allows us to fully determine τ_c and τ_e. Experimentally, however, we only see differences with respect to the equilibrium occupancy f^L. Thus, we have to consider

$$\Delta f(t_s) = f(t_s) - f^L = a(1 - e^{-t_s/\tau_c}),$$
$$\Delta f(t_s, t_r) = f(t_s, t_r) - f^L = a(1 - e^{-t_s/\tau_c})e^{-t_r/\tau_e},$$

with the maximum occupancy change given by $a = (f^H - f^L) \leqslant 1$. In the following we will use the normalized version $h(t_s, t_r; \tau_c, \tau_e) = \Delta f(t_s, t_r; \tau_c, \tau_e)/a$.

Provided that each defect k induces a shift in ΔV_{th} of $-\eta_k$, we obtain the overall degradation of N defects as

$$-\Delta V_{th}(t_s, t_r) = \sum_k^N \eta_k a_k h_k(t_s, t_r; \tau_{c,k}, \tau_{e,k}).$$

Note that just like the time constants, η_k is a stochastic quantity, and thus different for every defect. Most RTN experiments and simulation models claim that η_k is exponentially distributed [59,71,72], while a log-normal distribution has been suggested as well [73]. Recently, an exponential distribution was also observed in NBTI and PBTI experiments on small-area devices [74,56].

The maximum possible degradation $\Delta V_{th}^{max} = \Delta V_{th}(\infty, 0)$ is obtained when all defects have been eventually charged. In that case $h_k = 1$ and we get

$$\Delta V_{th}^{max} = \sum_k^N \eta_k a_k.$$

In general, for a defect with a single energy level E_1, we know that the equilibrium values f^L and f^H must be given by the Fermi–Dirac distribution,

$$f_{FD}(E_1) = \frac{1}{1 + e^{\beta E_{1F}}},$$

where $E_{1F} = E_1 - E_F$ is the distance of the defect level from the Fermi level and $\beta^{-1} = k_B T_L$. Since we are considering holes, we have to use $1 - f_{FD}(E_1)$. For simplicity we assume that no interaction with the gate occurs, which is roughly correct for defects located closer to the channel than to the gate, that is, $x \lesssim t_{ox}/2$. In this right half of the oxide, the equilibrium concentrations are determined by the Fermi-level of the channel and we have

$$f(E_1) = \frac{\tau_e}{\tau_e + \tau_c} = \frac{1}{1 + \tau_c/\tau_e} = 1 - f_{FD}(E_1).$$

Since the maximum possible amount of change in the occupancies is given by $a = f^H - f^L$, which depends on the bias conditions, the maximum possible degradation is bias dependent as well

$$\Delta V_{th}^{max} = \sum_k^N \eta_k \left(\frac{1}{1 + e^{-\beta E_{1F}^H}} - \frac{1}{1 + e^{-\beta E_{1F}^L}} \right).$$

This bias dependence comes simply from the fact that not all defects can contribute to ΔV_{th}. To first order, only defects in the active trapezoidal region where $f^H - f^L \approx 1$ can change their charge

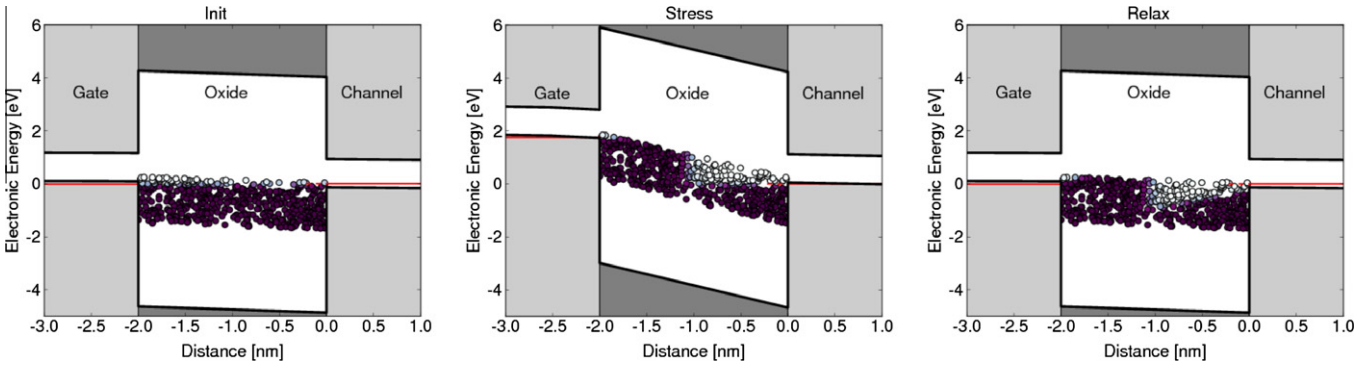

Fig. 32. Region of defects actively contributing to the degradation in a BTI setting. Filled circles are uncharged defects while the open circles symbolize the positively charged defects. **Left**: Prior to stress, it is assumed that all defects are in equilibrium with the Fermi-levels of the channel and the gate. Furthermore, the defect energy band is chosen in such a way that the contribution of defects located in right half of the oxide dominate the degradation. **Middle**: Following the switch to the stress voltage, a certain fraction of the defects is moved above the Fermi-level. As time progresses, only defects in the active region can contribute, which determines the maximum possible degradation. **Right**: Back at the recovery voltage, the defects are moved back below the Fermi-level.

state during stress, see Fig. 32. This corresponds to the region where prior to stress the defects are not occupied by a hole, $f^{\mathrm{L}} = 0$, but where the application of stress causes E_1 to be shifted above E_{F}. With E_1 above E_{F}, the defects will eventually be emptied, causing $f^{\mathrm{H}} = 1$. Defects with E_1 in the region below this trapezoid have $f^{\mathrm{H}} = 0$ as their trap-level will not be shifted above E_{F}. Consequently, we have there $a = 0$ and the defects do not contribute to the degradation. On the other hand, defects above the trapezoidal region are already occupied by a hole prior to stress, that is, $f^{\mathrm{L}} = 1$. As stress can therefore not discharge the defect any further, these defects cannot contribute either, signified by $a = 0$ as well.

4.1. Boundaries of the active region

The boundaries of the active region can be estimated by approximating the Fermi–Dirac distribution by a step function, $f_{\mathrm{FD}} \approx H(E_{\mathrm{F}} - E_1)$, where H is again the unit step function. Then we see that in order for $f^{\mathrm{L}} = 0$ to hold, E_1 of the defect must be below E_{F}. Also, the change in the bias conditions must move E_1 above E_{F} during stress. We thus have for the lower boundary $E_1^{\mathrm{H}} > E_{\mathrm{F}}$ and $E_1^{\mathrm{L}} < E_{\mathrm{F}}$ for the upper one.

The energy level E_1 at a certain position in the oxide depends on the bias via $E_1(x) = E_{10} - \mathrm{q}\varphi(x)$. Assuming to first-order that the charged defects inside the oxide do not significantly impact the electrostatic potential $\varphi(x)$, which is roughly valid only for low defect concentrations, we have $\varphi(x) = \varphi_{\mathrm{s}} - x\,F$, where φ_{s} is the potential at the interface and F the constant electric field across the oxide. The sign convention for F is such that application of a negative bias on the gate results in a positive F. Furthermore, x is the distance of the trap from the interface and positive. So in this simple approximation the trap level depends on the applied bias via $E_1 = E_{10} - \mathrm{q}\varphi_{\mathrm{s}} + \mathrm{q}x\,F$, see Fig. 33.

Thus, in terms of E_{10}, the conditions $E_1^{\mathrm{H}} > E_{\mathrm{F}}$ and $E_1^{\mathrm{L}} < E_{\mathrm{F}}$ transform to

$$E_{\mathrm{F}} + \mathrm{q}\varphi_{\mathrm{s}}^{\mathrm{H}} - \mathrm{q}xF^{\mathrm{H}} \lesssim E_{10}(x) \lesssim E_{\mathrm{F}} + \mathrm{q}\varphi_{\mathrm{s}}{}^{\mathrm{L}} - \mathrm{q}xF^{\mathrm{L}}.$$

The above defines the active region of defects which are neutral prior stress but can be charged for times larger than the longest time constant, see Fig. 34. Note that this consideration depends solely on equilibrium thermodynamic properties and is therefore valid for all physical models of single energy-level defects.

4.2. The capture/emission time map

A description employing individual defects is useful for small-area devices where the total number of defects is small. Large-area

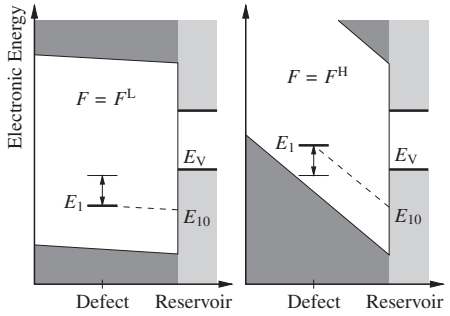

Fig. 33. Illustration of the impact of the electric field on the energy levels of an oxide defect. At small F, $E_1 < E_2 = E_{\mathrm{v}}$ (left). Application of a large F moves E_1 above E_2 (right).

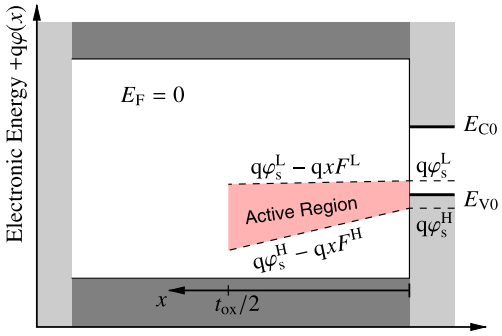

Fig. 34. The active region of BTI depends dominantly on the field during stress. Only defects in the active region are neutral prior stress and can be potentially charged during stress.

devices, however, have thousands of defects. In order to describe the overall degradation, it is impossible to consider each defect individually. Rather, similar defects can be grouped together using a suitably defined density [66], see Fig. 35. This density will be called *continuous CET map*, or simply CET map in the following. It is important to remember that in all the switched experiments discussed in the following, the capture time is recorded at the high-level while the emission time is taken at the low-level.

We formally define the discrete CET map g_{ij} as

$$g_{ij} = g(\tau_{\mathrm{c},i}, \tau_{\mathrm{e}j}) = \sum_k^N \frac{\eta_k a_k}{\Delta\tau^2} \mathrm{rect}\Big(\frac{\tau_{\mathrm{c},k} - \tau_{\mathrm{c},i}}{\Delta\tau}\Big) \mathrm{rect}\Big(\frac{\tau_{\mathrm{e},k} - \tau_{\mathrm{e}j}}{\Delta\tau}\Big),$$

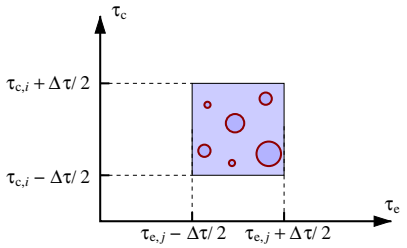

Fig. 35. Rather than considering a large number of individual defects, similar defects around $(\tau_{c,i}, \tau_{e,j})$ are grouped together in the density $g_{i,j}$. The size of the circle symbolizes the magnitude of the macroscopic step height, $\eta_k\, a_k$.

with the rectangle function

$$\mathrm{rect}(\tau/\Delta\tau) = \begin{cases} 1 & \tau - \Delta\tau/2 \leqslant \tau < \tau + \Delta\tau/2 \\ 0 & \text{otherwise} \end{cases}.$$

This definition simply collects all defects with $\tau_{c,k}$ and $\tau_{e,k}$ in the vicinity of $\tau_{c,i}$ and $\tau_{e,j}$ into $g_{i,j}$.

With $g_{i,j}$ at hand, we can now express the threshold voltage shift as

$$\Delta V_{\mathrm{th}}(t_s, t_r) \approx \sum_i \sum_j g(\tau_{c,i}, \tau_{e,j}) h(t_s, t_r; \tau_{c,i}, \tau_{e,j}).$$

In the limit $\Delta\tau \to 0$ we obtain the continuous CET map $g(\tau_c, \tau_e)$ as

$$g(\tau_c, \tau_e) = \sum_k^N \eta_k a_k \delta(\tau_{c,k} - \tau_c)\delta(\tau_{e,k} - \tau_e),$$

and eventually

$$\Delta V_{\mathrm{th}}(t_s, t_r) \approx \int_0^\infty \mathrm{d}\tau_c \int_0^\infty \mathrm{d}\tau_e g(\tau_c, \tau_e) h(t_s, t_r; \tau_c, \tau_e).$$

The main feature of h is that it selects all defects with $\tau_c < t_s$ and $\tau_e > t_r$, so to good approximation we can replace h by

$$h(t_s, t_r; \tau_c, \tau_e) \approx H(t_s - \tau_c) H(\tau_e - t_r).$$

Although this approximation is somewhat crude, as the two transitions contained in h cover a decade in time, it gives us a very simple and intuitive connection between ΔV_{th} and g,

$$\Delta V_{\mathrm{th}}(t_s, t_r) \approx \int_0^{t_s} \mathrm{d}\tau_c \int_{t_r}^\infty \mathrm{d}\tau_e g(\tau_c, \tau_e). \tag{47}$$

In words this means that ΔV_{th} is given by the sum of all defects charged until t_s but not yet discharged after t_r. Eq. (47) can now be used to give a simple method for the extraction of g by simply taking the negative mixed partial derivative of a given ΔV_{th} stress/recovery data set,

$$g(\tau_c, \tau_e) \approx -\frac{\partial^2 \Delta V_{\mathrm{th}}(\tau_c, \tau_e)}{\partial\tau_c \partial\tau_e}. \tag{48}$$

If the CET map is represented on logarithmic axes, one has to use

$$\tilde{g}(\tau_c, \tau_e) = \tau_c \tau_e g(\tau_c, \tau_e), \tag{49}$$

in analogy to (14). While g gives the density of defects per unit time, for example information on how much ΔV_{th} is gained/lost in a second, $\tilde{g}$ gives the density on a logarithmic scale, for example on how much ΔV_{th} is gained/lost per decade.

A typical experimental CET map obtained in that manner is shown in Fig. 36. Analysis of the CET map shows that for a fixed τ_c, the distribution of τ_e is roughly Gaussian. Furthermore, τ_e and τ_c appear correlated, with both distributions being rather wide [66,75].

Finally, it is intuitive to consider the fraction of the CET map visible in experimental data. In delay-free experiments, which have become known as on-the-fly (OTF) measurements, it is attempted

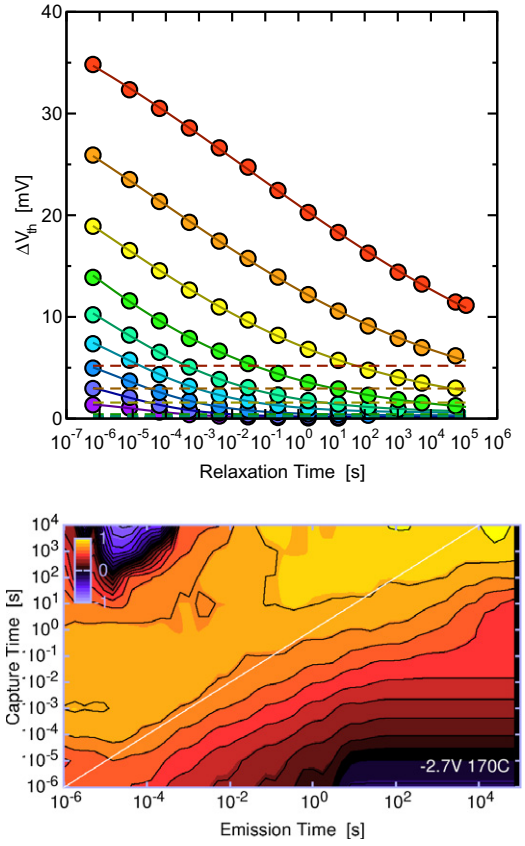

Fig. 36. Using experimental ΔV_{th} recovery traces with an as wide as possible window in both t_s and t_r (top), the CET map can be constructed by taking the mixed partial derivative (bottom) [66]. The solid lines in the top figure are obtained by integrating the CET map following (47). The dashed lines are permanent component not visible in the CET map [75].

to record the NBTI degradation directly under stress conditions [76]. As such, the recorded degradation of $I_D(t)$ has to be converted to a threshold voltage shift. Since the degradation is monitored at V_G^H without switching to V_G^L, the measurement delay is zero and the CET map covers the whole τ_e axis. In practice, however, a delay-free experiment requires determination of a reference value of $I_D(V_G^H)$ for the calculation of ΔV_{th} [77,78]. This reference value is determined with a certain delay t_M. As a result, even OTF measurements do not capture all defects as the lower part of the τ_c axis with $\tau_c < t_M$ is missed.

The dual problem is observed for conventional measure–stress–measure (MSM) setups [32,79]. In an MSM measurement, the reference value of $I_D\left(V_G^L\right)$ is determined first. Then, the device is stressed for a duration of t_s. After termination of the stress, $I_D\left(V_G^L\right)$ is continuously measured as quickly as possible with a minimum delay of t_M. Thereby, the MSM sequence covers the complete τ_c axis up to $\tau_c < t_s$ but only a part of the τ_e axis where $\tau_e > t_M$. The difference between the two setups is visualized in Fig. 37. As such, neither method is able to "see" all defects present in the CET map [79].

4.3. Theoretical CET maps

A number of theoretical or empirical models for $\Delta V_{\mathrm{th}}(t_s, t_r)$ have been published either in the context of NBTI or charge trapping. Using (48), we will derive the resulting CET map for a few important cases: first, the classic reaction–diffusion model for NBTI, then a standard hole trapping model, and finally the empirically found universal recovery relationship.

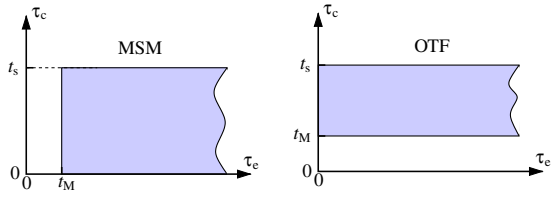

Fig. 37. Regions of the CET map scanned by MSM (left) versus OTF setups (right). Neither measurement captures all defects: The MSM setup misses all defects with $\tau_e < t_M$ while the OTF setup misses all defects with $\tau_c < t_M$.

4.3.1. The reaction–diffusion model

The most popularized NBTI model is based on the reaction–diffusion (RD) formalism originally suggested by Jeppson and Svensson [10]. Although the RD model contains a reaction step, the breaking of Si–H bonds at the interface, this reaction is assumed to be in quasi-equilibrium. The overall degradation is assumed to be dominated by the diffusion profile of hydrogen inside the oxide. The model has seen a number of refinements during the last couple of years [18,80,81], with the current version relying on the diffusion of neutral H_2. The most recent extension postulates the existence of an elastic hole trapping component, which obscures both the stress and recovery data for the first second [82,83]. In this version of RD theory, hole trapping is considered a mere experimental nuisance which has to be removed from the data in order to unveil the true, diffusion-limited, NBTI degradation [30].

As hole trapping and hydrogen diffusion are considered to be independent, the CET maps can be constructed independently for both mechanisms. The total CET map is then obtained by adding the two maps. We thus defer the discussion of hole trapping for the time being to Section 4.3.2 and begin with the diffusion-limited regime of the RD model. The following simple analytical expression gives excellent agreement with the numerical solution of the RD model [23]

$$\Delta V_{\mathrm{th}}(t_s, t_r) = \frac{t_s^n}{1+\sqrt{t_r/t_s}}, \tag{50}$$

where $n = 1/6$. Saturation at large t_s is neglected for simplicity as it does not change the basic picture. The normalized recovery predicted by the RD model is universal in the sense that it depends only on the ratio of t_r/t_s [23]. Using (49), we directly calculate the logarithmic CET map $\tilde{g}$ from (50) as

$$\tilde{g}(\tau_c, \tau_e) = \frac{\sqrt{\tau_e \tau_c}(2n-1) + \tau_e(2n+1)}{4\left(1+\sqrt{\tau_e/\tau_c}\right)^3} \frac{1}{\tau_c^{1-n}}. \tag{51}$$

The logarithmic CET map of the RD model is shown in Fig. 38. It bears only minimal resemblance to the experimentally obtained CET shown in Fig. 36. Remarkably, $\tilde{g}$ becomes negative for $\tau_e < \tau_c/4$. Physically, this means that "time constants" available at the certain stress time are no longer available after longer stress times. The reason for this is that due to the diffusive nature of the model, recovery takes longer with increasing stress times, as the H_2 molecules have diffused away into the oxide. Recovery requires them to come back, which takes increasingly longer with increasing t_s, see Fig. 39. Such a behavior is in contrast to the experimental data, which show emission times independent of stress times [65,66,3].

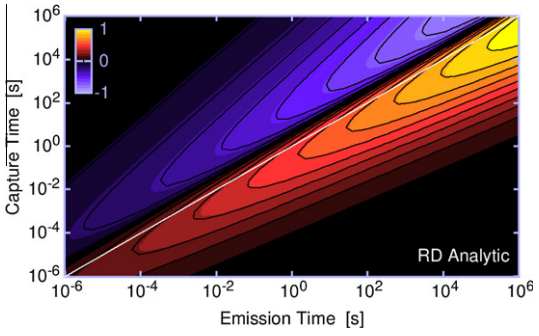

Fig. 38. Logarithmic CET map predicted by the RD model. Note that the entries in the map above the line $\tau_e = \tau_c/4$ are negative.

4.3.2. Simple hole trapping

In its simplest form, hole trapping into the oxide is considered to be a purely elastic process describable in an extended version of the SRH model [84,21,82,83], see Section 5. An analytical solution for this scenario has been developed by Tewksbury [85]

$$\Delta V_{\mathrm{th}}(t_s, t_r) = A \log(1 + B t_r/t_s) \quad \text{for } t_s < t_s^{\max}, \tag{52}$$

with the factor B accounting for a possible difference in $\bar{\tau}_c$ and $\bar{\tau}_e$ directly at the interface. In this elastic model, the capture and emission times are solely determined by the depth of the defect into the oxide, which leads to saturation of ΔV_{th} once the deepest defect has been charged at $t_s = t_s^{\max}$. Since modern oxides are very thin, for instance 2nm, this saturation occurs relatively early, for instance at about $t_s^{\max} = 1$ s.

The logarithmic CET map is obtained from the above as

$$\tilde{g}(\tau_c, \tau_e) = \frac{AB}{(B+\tau_e/\tau_c)^2} \frac{\tau_e}{\tau_c} \quad \text{for } \tau_c < t_{\max}. \tag{53}$$

The logarithmic CET map of this simple hole trapping model is a logistic distribution of the ratio τ_e/τ_c. In its pseudo two-dimensional form it is shown in Fig. 40 for two values of B. With increasing τ_c, the one-dimensional logistic distribution is shifted towards larger emission times until $\tau_c = t_{\max}$.

4.3.3. Universal recovery

It has been observed that recovery following NBTI/PBTI stress in both nMOS and pMOS transistors can be empirically described by [33]

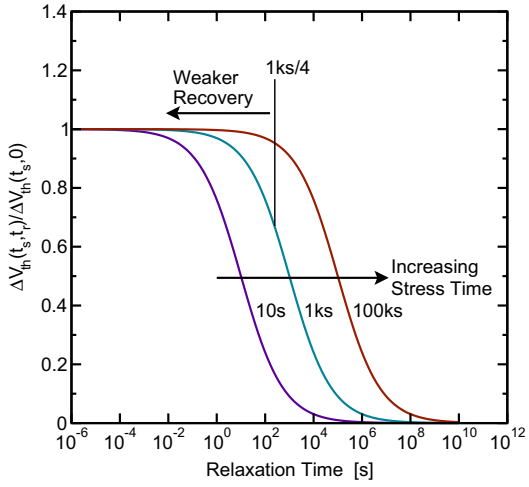

Fig. 39. Schematics showing why the RD model generates negative entries on the CET map. The plot shows the normalized recovery predicted by the RD model for three stress times. As can be seen, the shape of the curve does not change with time but is only shifted to larger times, with 50% recovery corresponding to $t_r = t_s$. Thus, with increasing stress time, recovery sets in later. As a consequence, recovery at $t_r < t_s/4$ is weaker than at the previous stress time. In order to reflect this delay of recovery in the CET map, negative entries are required for $t_r < t_s/4$.

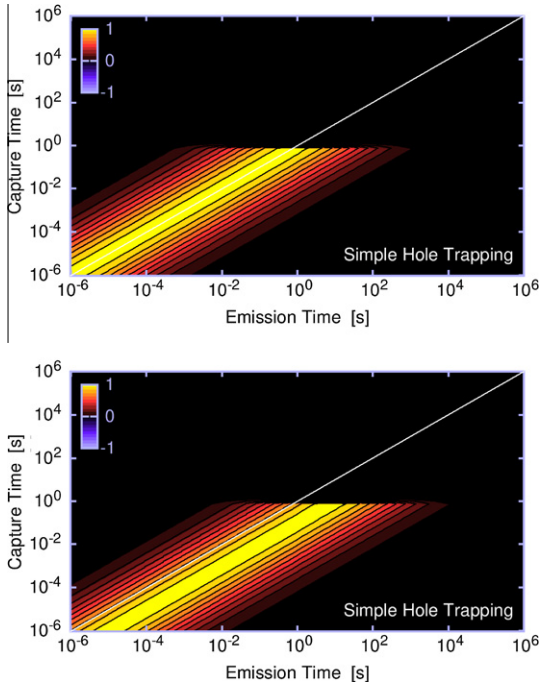

Fig. 40. The CET which generates the simple hole trapping model of (52). **Left**: $B = 1$ and **Right**: $B = 10$, which simply corresponds to a shift to larger emission times by a factor of 10.

$$\Delta V_{\mathrm{th}}(t_s, t_r) = \frac{A t_s^a}{1 + B(t_r/t_s)^b} + P t_s^n. \tag{54}$$

In this empirical model, the first term describes the recoverable component while the second term forms a permanent contribution, which, as such, does not contribute to the CET map. Note that the universal relationship (54) reduces to the RD model with the

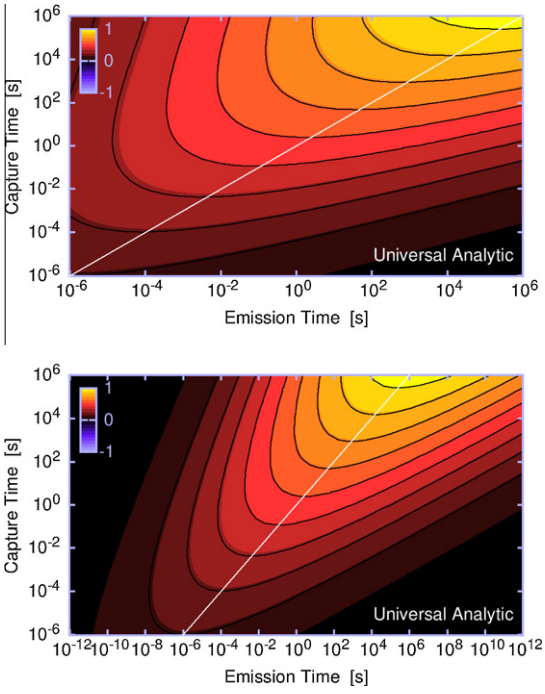

Fig. 41. The CET map of the universal relationship. **Top**: A typical parameter combination consistent with experimental data, $a = 1/6$, $b = 0.15$, $B = 2$. **Bottom**: Same parameters as above but with $b = a$ and plotted over a wider range.

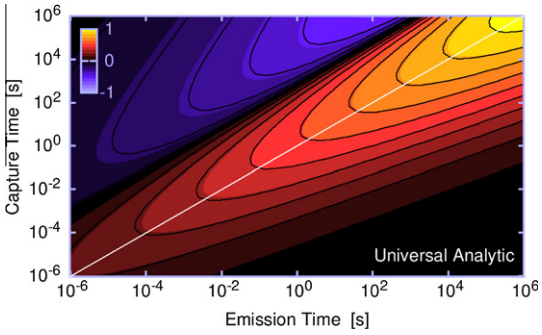

Fig. 42. The CET of the universal relationship has negative entries when $b > a$. $a = 1/6$, $b = 0.3$, $B = 2$.

parameters $B = 1$, $a = 1/6$, $b = 1/2$, and $P = 0$. We again use (49) to directly calculate $\tilde{g}$ from ΔV_{th} as

$$\tilde{g}(\tau_c, \tau_e) = bAB \frac{(a-b)(\tau_e \tau_c)^b + (a+b) B \tau_e^{2b}}{(1 + B(\tau_e/\tau_c)^b)^3} \frac{1}{\tau_c^{2b-a}}.$$

As with the RD model, the CET map will have negative entries when $a < b$ and

$$\tau_e < \tau_c \left(B \frac{b+a}{b-a} \right)^{1/b}.$$

Experimentally, b was observed to be in the range $[0.1 \ldots 0.2]$ [23,86,33], which is close to the typical NBTI power-law slope of $a \approx 0.15$, so the special case $b \approx a$ is of interest. There we get

$$\tilde{g}(\tau_c, \tau_e) = aAB \frac{2aB\tau_e^{2a}}{(1 + B(\tau_e/\tau_c)^a)^3} \frac{1}{\tau_c^a}.$$

The CET map of the universal relationship is shown in Figs. 41 and 42 for a few parameter combinations.

5. SRH-like models for the transition rates

In the following we will discuss physical models for the rates describing the transition of carriers between defects and a reservoir. Typical reservoirs are the channel and the gate of an MOS transistor. A rather general framework for the calculation of such transition rates was given by Shockley and Read in their famous paper [43] for what has become known as the Shockley–Read–Hall (SRH) recombination process. Although the SRH model was originally derived for recombination centers located in the bulk of a semiconductor where the defect and the carrier reservoirs are located at the same position, the model has been generalized to account for trapping into oxides. We shall start with the same formalism and proceed by discussing two important models for the actual transition rates. The first model is a generalized version of the SRH process which only considers electronic energies. Irrespective of the fact that this is the most commonly used approach, it must be clearly stated that this model is only valid for a certain class of bulk defects but cannot describe oxide defects. The failure of the SRH model to describe oxide defects has been known for a long time [48,49], but it will quite obviously require continued efforts [51] until the majority of researchers will refrain from using it.

In the following we calculate the rates for a single donor-like defect, assuming only interactions with the valence band. The defect will be assumed to be in one of its two states, 1 for neutral and 2 for positive. The expectation values of the defect to be in either state are f_1 and f_2, with $f_1 + f_2 = 1$. We consider a system consisting of a defect plus an electron, which can be moved back and

forth between the defect and the reservoir. When the electron is at the defect site, its energy is E_1, when it is moved to the reservoir the energy changes to E_2. The differential transition rates of the two partial processes are then

$$\mathrm{d}k_{12}(E) = c_p(E) f_p(E) g_p(E)\,\mathrm{d}E, \quad (f_1), \tag{55}$$

$$\mathrm{d}k_{21}(E) = e_p(E) f_n(E) g_p(E)\,\mathrm{d}E, \quad (f_2), \tag{56}$$

where the term in parenthesis gives the state the defect has to be in for the rate to apply. The electron energy distribution function in the reservoir is f_n, the hole distribution function $f_p = 1 - f_n$, while g_p is the density-of-states in the valence band. The physics of the capture and emission process go into the energy-dependent capture and emission coefficients c_p, and e_p, which will receive our full attention later.

To obtain the transition rates, we have to integrate Eqs. (55) and (56) over all possible states in the valence band. Unfortunately, the resulting equations cannot be further processed analytically without making assumptions on the form of f_n and the two capture/emission coefficients. With regard to the distribution function, we will assume that f_n is given by the Fermi–Dirac distribution $f_{\mathrm{FD}}(E)$. This is exact in thermal equilibrium and a very good approximation for bias temperature stress and recovery experiments, because there the current flow in the channel is usually negligible [87,88]. The Fermi–Dirac distribution has the useful property

$$\frac{f_{\mathrm{FD}}(E)}{1 - f_{\mathrm{FD}}(E)} = \mathrm{e}^{-\beta(E - E_{\mathrm{F}})}. \tag{57}$$

We furthermore introduce energy averages of a quantity $h(E)$ over the valence band

$$\langle\!\langle h(E) \rangle\!\rangle = \frac{1}{p} \int_{-\infty}^{E_{\mathrm{v}}} h(E) f_p(E) g_p(E)\,\mathrm{d}E. \tag{58}$$

The energy average is defined in such a way that for energy-independent quantities h one obtains $\langle\!\langle h \rangle\!\rangle = h$.

Substituting the Fermi–Dirac distribution and integrating over the valence band, one obtains

$$k_{12} = \int_{-\infty}^{E_{\mathrm{v}}} c_p(E) f_p(E) g_p(E)\,\mathrm{d}E = p \langle\!\langle c_p(E) \rangle\!\rangle, \tag{59}$$

$$k_{21} = \int_{-\infty}^{E_{\mathrm{v}}} e_p(E) \mathrm{e}^{-\beta(E - E_{\mathrm{F}})} f_p(E) g_p(E)\,\mathrm{d}E = p \langle\!\langle e_p(E) \mathrm{e}^{-\beta(E - E_{\mathrm{F}})} \rangle\!\rangle. \tag{60}$$

Overall, with the above given rates, the temporal change of f_1 is then given as in (5) by

$$\frac{\partial f_1}{\partial t} = f_2 k_{21} - f_1 k_{12}.$$

The rates derived so far are of limited practical use since they still require physical models for the capture and emission coefficients, c_p and e_p, which have to be temperature- and bias-dependent in order to correctly reflect the experimental behavior.

5.1. The SRH model

In the literature, the transition between the states of a defect are predominantly described using only their electronic energy levels, as is done for instance in conventional SRH theory. Despite the fact that the application of the SRH model to oxide defects has raised many doubts as to its theoretical justification [49,85], it is almost exclusively used by reliability engineers for the analysis of charge trapping events. As a number of more or less elaborate derivations are available in literature, see for instance Refs. [46,84,85], we will restrict ourselves to a qualitative derivation which captures the essential features of the model.

For instance, consider a defect which can either contain an electron or not. Such a defect can be either an electron trap (negatively charged after capture of the electron) or a hole trap (positively charged after emission of the electron). In the electronic-energy-only picture, the energy of the defect is then solely determined by the potential energy of that electron. For instance, a donor-like defect is electrically neutral when the electron is located at the defect site at the energy E_1. Conversely, the defect can emit this electron which is then moved to the reservoir, which could be the valence band in the channel. This emitted electron is then located in the reservoir, which in a simplest approximation is considered to only have a single energy level, $E_2 = E_{\mathrm{V}}$. The two energy levels E_1 and E_2 give the energy of the defect when in state f_1 or f_2, respectively.

With regard to the standard SRH model, actually no explicit physical model was invoked for the capture and emission coefficients in the original paper. The implicit assumption was that hole capture from the valence band occurs without a barrier and is simply proportional to a capture cross section times the thermal velocity of the carriers,

$$c_p(E) = \nu_{\mathrm{th}} \sigma, \tag{61}$$

with the thermal velocity $\nu_{\mathrm{th}} = \sqrt{8 \mathrm{k_B} T_{\mathrm{L}}/(\pi m)}$ [50].

While the traditional SRH model was derived for bulk defects, it has been used extensively to also describe charge exchange with oxide defects [46,84,44]. In order to consider the different spatial location of the reservoir and the defect, it is assumed that the carriers tunnel elastically between the defect and the reservoir. Tunneling is considered in a more-or-less empirical manner by multiplying the capture cross-section with the tunneling coefficient ϑ calculated in the WKB approximation as

$$\vartheta = \exp\left[-\frac{4\sqrt{2m}}{3\hbar \mathrm{q} F} \left((\mathrm{q}\phi - E)^{3/2} - (\mathrm{q}\phi_0 - E)^{3/2} \right) \right], \tag{62}$$

with m as the tunneling mass, a somewhat ill-defined fit parameter on the order of the free electron mass, ϕ the barrier at the interface and ϕ_0 the barrier at the defect site.

Since in many cases defects close to the interface are considered where $x < 1$ nm, the barrier is very thin and very high. As such, the modulation of the barrier by the local potential can be neglected and we obtain via a Taylor expansion of the $(\cdot)^{3/2}$ terms

$$\vartheta \approx \mathrm{e}^{-x/x_0}, \tag{63}$$

with $x_0 = \hbar / \left(2\sqrt{2m\phi} \right)$. A comparison of the full WKB model (62) with the simplified version (63) is given in Fig. 43. This approximation is useful for analytical estimates of the field-dependence, since only for traps deeper into the oxide the field-dependence of ϑ becomes relevant.

It is worth mentioning that simply multiplying the capture cross section with the WKB coefficient is a rather crude approximation of the real physical problem. However, more accurate attempts, where the transition matrix element is evaluated, require knowledge of the defect potential, which is essentially unknown. Usually employed assumptions like δ function potentials lead to very similar results as those obtained from the WKB approximation [89,85]. Given the degree of uncertainty regarding the defect potentials, we consider the WKB approximation sufficient for our present purposes. The tunneling coefficient ϑ is then usually pulled into the capture cross section, which gives us

$$\sigma = \sigma_0 \vartheta. \tag{64}$$

Also, for oxide defects, the thermal velocity should be replaced by a quantity which only considers carriers with a velocity component normal to the interface [85]. However, this is believed to be a second-order effect and will not be considered in the following.

Due to the requirement of detailed balance, the following must hold in thermal equilibrium for any energy E,

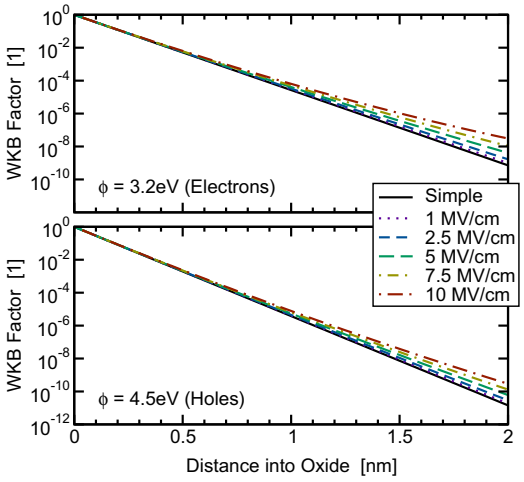

Fig. 43. The WKB factor for electrons (top) and holes (bottom) as a function of the defect location inside the oxide. Parameter is the electric field, E is assumed to be zero.

$$f_1\,\mathrm{d}k_{12}(E) = f_2\,\mathrm{d}k_{21}(E). \tag{65}$$

From the above together with (59) and (60) we obtain the relation

$$\frac{c_p(E)}{e_p(E)} = \frac{f_2}{f_1}\frac{f_n(E)}{f_p(E)} = \frac{1-f_1}{f_1}\frac{f_n(E)}{1-f_n(E)}, \tag{66}$$

and with $f_1^0 = f$ and $f_n^0 = f$ finally

$$\frac{c_p^0(E)}{e_p^0(E)} = \mathrm{e}^{-\beta(E_1-E_\mathrm{F})}\mathrm{e}^{\beta(E-E_\mathrm{F})} = \mathrm{e}^{-\beta(E-E_1)}. \tag{67}$$

The above relates capture and emission coefficients by a Boltzmann factor. The energy level of the defect in state 1 is given by E_1, which is conventionally denoted by E_T in the literature. However, as we shall be shortly dealing with defects having more than one trap level, we use this slightly more general notation. The fundamental assumption is now that relationship (67) is valid also under non-equilibrium conditions.

In order to explicitly calculate the emission coefficients we assume Boltzmann statistics to hold, that is, $p = N_\mathrm{V}\exp(\beta E_\mathrm{VF})$. The notation E_{12} is a shorthand for $E_1 - E_2$ and will be frequently used in the following. Using (67), we obtain

$$k_{21} = \langle\!\langle c_p(E)\rangle\!\rangle N_\mathrm{V}\mathrm{e}^{\beta E_{\mathrm{V}1}}.$$

Using the energy-independent capture coefficients of (61), the transition rates of the SRH model are

$$k_{12} = p\nu_\mathrm{th}\sigma, \quad k_{21} = N_\mathrm{V}\nu_\mathrm{th}\sigma\mathrm{e}^{\beta E_{\mathrm{V}1}}. \tag{68}$$

Frequently, the above rates are used irrespective of the energetic position of the defect relative to the reservoir. This is incorrect, as will be discussed now.

5.2. The SRH model for states outside the reservoir bandgap

In the standard SRH model for semiconductor bulk defects, the defect energy level is reasonably assumed to be inside the bandgap of the material, $E_\mathrm{V} < E_1 < E_\mathrm{C}$. Then, the Boltzmann factor in k_{12} is smaller than unity. Contrary to semiconductor bulk defects, oxide defects can very well have their energy-level outside the bandgap of the reservoir, making the Boltzmann factors larger than unity. The reason for this is that in the case of defects outside the silicon bandgap the assumption that electron capture from the conduction band and hole capture from the valence band proceed without barrier is certainly wrong, as the electron has to be raised to an energy level higher than the reservoir while the hole would have to be pushed to an energy level lower than the reservoir.

So in analogy to (61), it appears now more prudent to assume the emission coefficients to be energy independent which gives

$$k_{12} = pe_p\langle\!\langle \mathrm{e}^{-\beta(E-E_1)}\rangle\!\rangle, \quad k_{21} = pe_p\langle\!\langle \mathrm{e}^{-\beta(E-E_\mathrm{F})}\rangle\!\rangle. \tag{69}$$

Unfortunately, the above cannot be simplified with the same elegance as if the defect level were inside the bandgap. One reason for this is that during the integration over all bandstates, at one point the energy E will become higher than the trap level E_1. Then, the capture coefficient should be energy-independent rather than the emission coefficient. Overall, even for parabolic bands and a Boltzmann distribution rather complicated expressions are obtained. We will therefore make a very crude approximation by assuming that $E = E_\mathrm{V}$ for the calculation of k. Together with the Boltzmann distribution and $e_p = \nu_\mathrm{th}\sigma$ one then obtains

$$k_{12} = p\nu_\mathrm{th}\sigma\mathrm{e}^{-\beta E_{\mathrm{V}1}}, \quad k_{21} = N_\mathrm{V}\nu_\mathrm{th}\sigma. \tag{70}$$

The above looks very similar to (68), with the exception that the inverse of the Boltzmann factor now appears with the rate k_{12} rather than with k_{21} [44]. This makes sense intuitively, since now hole capture into the defect level $E_1 < E_\mathrm{V}$ requires the thermal activation of holes. Still, it must be repeated that this derivation is extremely crude. However, considering the other shortcomings of the SRH model with respect to oxide defects, we shall not be bothered unduly by this.

5.3. Heuristic interpretation of the SRH model

As simple but somewhat heuristic way to derive the rates of the SRH model without resorting to the use of the detailed balance relation (65) is by expressing the different occupation probabilities of the various energy levels directly using Boltzmann factors. Naturally, both methods will give similar results, as they follow directly from the same thermodynamic arguments.

In order to properly determine the energetic barrier for the transition, the two cases shown in Fig. 44 have to be considered: first, in the case the $E_1 < E_2$, the energy of the electron has to be increased in order to allow for the transition. The energy required is supplied by the reservoir which is assumed to be at temperature T. The probability that an electron with energy $\mathcal{E} = E_1$ is raised to E_2 is given by the Boltzmann factor

$$P\{\mathcal{E} = E_2|\mathcal{E} = E_1\} = \frac{\mathrm{e}^{-\beta E_2}}{\mathrm{e}^{-\beta E_1}} = \mathrm{e}^{-\beta E_{21}}, \tag{71}$$

with $E_{21} = E_2 - E_1$. On the other hand, there is no energetic barrier for the transition from state 2 to state 1 because $E_1 < E_2$.

In the second case shown in Fig. 44 we have $E_1 > E_2$. Then there is now barrier for the transition from state 2 to state 1. Conversely, the barrier $E_{12} = -E_{21}$ has to be surmounted for the transition from state 2 to state 1.

Both cases can be combined and we write the barrier for the transitions $1 \to 2$ and $2 \to 1$ as

$$\mathcal{E}_{12} = \max(E_{21}, 0) \quad \text{and} \quad \mathcal{E}_{21} = \max(E_{12}, 0). \tag{72}$$

With these barriers, the transition rates can be expressed in a more general way as

$$k_{12} = p\nu_\mathrm{th}\sigma\mathrm{e}^{\beta\mathcal{E}_{12}}, \quad k_{21} = N_\mathrm{V}\nu_\mathrm{th}\sigma\mathrm{e}^{\beta\mathcal{E}_{21}}. \tag{73}$$

5.4. Field and temperature dependence of the SRH model

Three terms contribute to the field dependence of the SRH model shown in Fig. 45. First, the exponential contribution due to the thermal emission barriers $\mathcal{E}_{12}$ and $\mathcal{E}_{21}$ which either contribute to

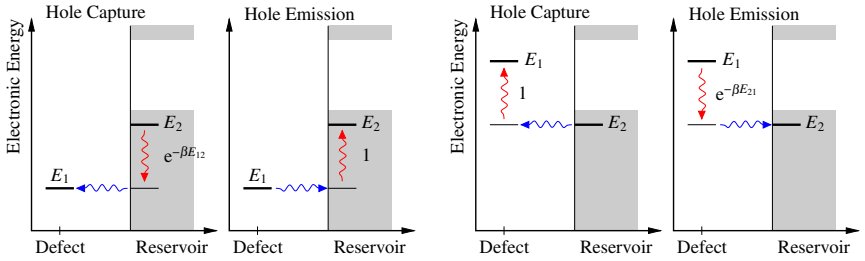

Fig. 44. The energetic barriers encountered in a model which considers only the electronic energies of the defect to describe a hole capture process. The barrier is determined by the energy difference $E_{21} = E_2 - E_1$ which can be positive (left) or negative (right). **Left**: When $E_2 > E_1$, phonons have to be absorbed to raise the hole from E_2 to E_1 in order to allow capture through an elastic tunneling process. During emission, phonons are emitted following the tunneling process. **Right**: When $E_2 < E_1$, no phonons have to be absorbed during capture. However, emission of the hole is only possible after sufficient energy has been absorbed from phonons.

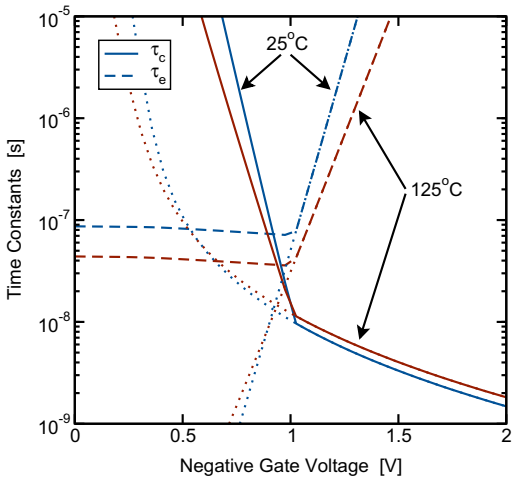

Fig. 45. The field and temperature dependence of capture and emission times obtained from the SRH model, evaluated for pMOS transistor with a 2 nm thick oxide at two temperatures. The defect is located at $x = 1$ nm and 0.4 eV below E_V. At about $|V_G| = 1$ V the defect moves into the silicon bandgap, visible by a kink in both τ_e and τ_c. The dashed line corresponds to the case when the defect is always considered to be inside the bandgap. Compared to experimental data, the time constants are too small (nanosecond to millisecond regime compared to experimental values well above kiloseconds). Furthermore, at high $|V_G|$, τ_c becomes larger with increasing temperature, which is in contrast to the strong decrease observed experimentally.

τ_c or τ_e, depending on the energetic position of the trap level E_1 relative to E_2. Since both levels shift with the surface potential, the barrier E_{21} is independent of φ_s but to first-order decreases linearly with increasing oxide field F as

$$E_{21} = E_{20} - E_{10} - qxF. \tag{74}$$

Second, the surface hole concentration gives a strong contribution below the threshold voltage which becomes considerably weaker at higher stress voltages. Finally, the WKB factor causes the weakest bias dependence, notable mostly in τ_e at low V_G.

The dominant temperature dependence stems from the thermal emission barriers. Since they are only relevant for τ_c when $E_1 < E_2$ and for τ_e when $E_1 > E_2$, the temperature dependence predicted by the SRH model is quite different from experimental observations, see Figs. 28 and 29. In particular, at large $|V_G|$ the temperature dependence of τ_c is only due to the temperature dependence of p, resulting in an *increase* of τ_c with increasing T.

5.5. Time dependence during stress and recovery

The time dependence of the defect occupancies during a BTI experiment as predicted by the SRH model is shown in Fig. 46. We have seen before that only defects inside the active region can contribute to ΔV_{th}. This means that during stress only defects above E_F are relevant. Although E_F will be a little bit below E_V during stress, resulting in a small thermal barrier $\mathcal{E}_{12}$ during capture, most defects will be above E_V, having no thermal capture barrier at all. Analogously, during recovery, these defects will be moved below E_F. Since then usually $E_F \gtrsim E_V$, most defects will be able to emit their captured hole with only a small barrier. So, to first order, the general formulas (73) simplify to

$$k_{12} \approx p\nu_{th}\sigma, \quad k_{21} \approx N_V\nu_{th}\sigma. \tag{75}$$

From this we see that both capture and emission are independent of the trap level E_1, which only determines whether the defect is inside the active area or not. As a consequence, the time constants will be

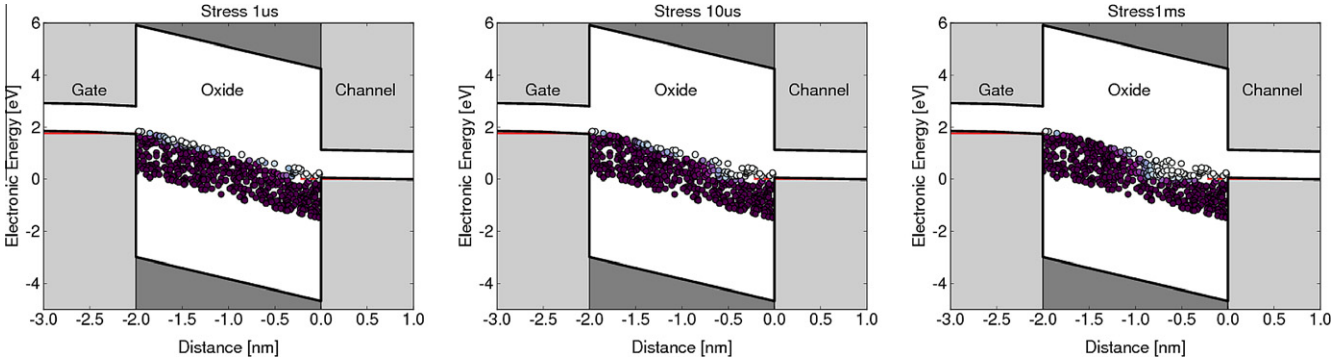

Fig. 46. Time dependence of the charging/discharging transients according to the SRH model. Prior to stress, it is assumed that all defects are in equilibrium with the Fermi-levels of the channel and the gate, see Fig. 32. With increasing stress time (from left to right), a tunneling front progresses from the interface towards the middle of the device. At 1 ms, all defects inside the active region are charged and the tunneling front stops. The situation is analogous during recovery, where the defects are neutralized in a tunneling front progressing from the right towards the center.

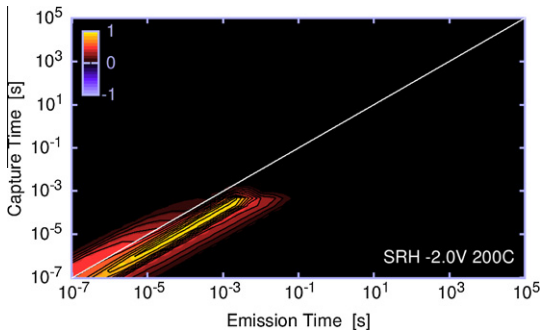

Fig. 47. The CET map of the SRH model, which is essentially a narrow stripe around $\tau_c = \tau_e$.

only weakly temperature and bias dependent. Furthermore, the time constants are only determined by the depth of the trap into the oxide, x, which enters both time constants in the same manner via the WKB factor incorporated in the capture cross section σ.

It is worthwhile to explicitly write down the SRH capture and emission times in their full glory

$$\tau_c \approx \frac{e^{x/x_0}}{p\nu_{th}\sigma_0}, \quad \tau_e \approx \frac{e^{x/x_0}}{N_V\nu_{th}\sigma_0}. \tag{76}$$

So even if we assume that an ensemble of defects with a wide distribution of E_1 and x exists, only the x distribution will enter the time constants. In particular, for a given x, the capture and emission times will be correlated. As a consequence, the CET map predicted by the SRH model, see Fig. 47, bears little resemblance with the experimental map of Fig. 36. We therefore must conclude that the SRH model is unable to describe oxide defects.

6. Nonradiative multiphonon transition rates

The reason for the failure of the SRH model is that it ignores the deformation of the defect site when the charge state is changed. Although the exact microscopic nature of the charge trapping sites is still not unequivocally established, the oxygen vacancy/E' center is the most commonly studied defect in silica and has been linked to BTI in a number of studies [91,13,92]. As an example, two charge states of the oxygen vacancy/E' center as calculated by density-functional-theory (DFT) are shown in Fig. 48. In the neutral equilibrium configuration, we have an oxygen vacancy, which is a silicon–silicon bond inside SiO_2. The distance between the two silicon atoms is larger than in crystalline silicon. Upon positively charging the oxygen vacancy, the distance between the silicon atoms increases even further, creating what has become known as an E' center. The naming convention is due to traditional electron spin resonance (ESR) analysis [6]. Eventually, one of the silicon atoms may move through the plane spanned by the three oxygen atoms into the so-called puckered configuration (an E'_γ center), thereby forming a bond with the oxygen atom behind it (not shown). Whether or not such a bond can be formed depends in an amorphous oxide on the availability of a suitable oxygen atom [93].

In the simplest picture, the equilibrium positions in either charge state are determined by the equilibrium between quantum–mechanical repulsive and attractive forces [94]. However, with increasing temperature, the atoms vibrate more and more vigorously around their equilibrium position in a chaotic fashion. Every displacement from the equilibrium position increases the total energy of the system. Naturally, the real motion of the atoms is highly complex and – particularly in the amorphous oxide we are dealing with – impossible to model precisely. As such, one usually limits oneself to the movement of the system along a dominant *reaction coordinate*. In general, the total energy along the reaction coordinate will have a complicated shape. Still, for reasonably small displacements from the equilibrium position, any energy–position relationship can be approximated by the lowest term of its Taylor expansion, which is a quadratic function of the displacement. The motion in this simplified parabolic potential is *harmonic* and model systems employing such harmonic oscillators are ubiquitous in solid-state physics. Since we are frequently dealing with rather strong distortions of the atomic positions, the "small displacement approximation" may appear unjustified. Nonetheless, such a simple model seems to be able to capture the essence of the phenomenon [95].

In the simplest case we again consider a two-state defect which is neutral in state 1 and positively charged in state 2. In each state the atomic equilibrium configuration is different, implying that the equilibrium position will depend on the defect state and will be denoted by q_1 and q_2. Also, in each state the total energy consists of contributions from the ionic system, the electronic system, and a coupling term. The coupling term is at the heart of the model and is responsible for the shift in the equilibrium positions as well as a change of the vibrational frequencies and will be discussed below. Overall, the total energy of each charge state i is usually written as

$$V_i = \frac{1}{2}M\omega_i^2(q-q_i)^2 + E_i, \tag{77}$$

where q is the reaction coordinate with the local equilibrium position q_i, M the effective mass of the 'defect molecule' [90], ω_i the vibrational frequency in minimum i, and E_i the potential energy in the minimum. Such a parabolic approximation to the total energies estimated from DFT calculations of the two charge states are also shown in Fig. 48 (middle) [90] and schematically in Fig. 49.

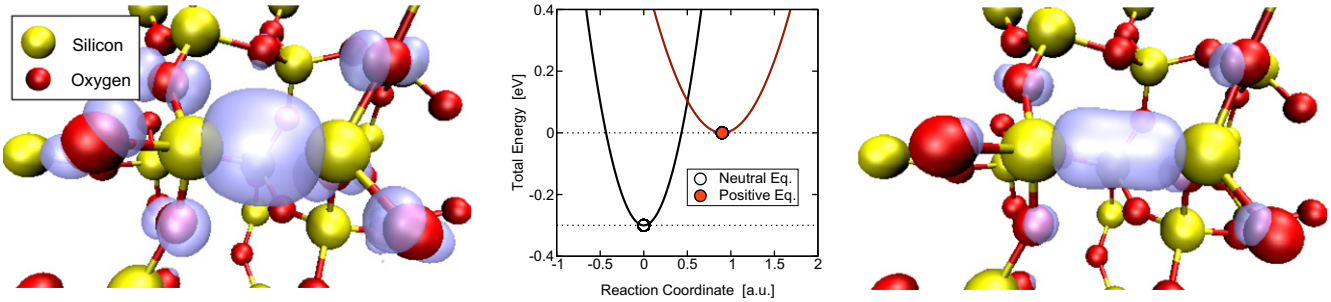

Fig. 48. Two charge states of the E' center calculated by density-functional-theory (DFT) [90], neutral (left) and positive (right). The electron density of the localized Kohn–Sham-eigenstate is shown as blue 'bubbles'. Note that the atomic equilibrium positions change when the charge state is changed. In the middle the total energy of the two states is given as a function of the reaction coordinate (dominant path).

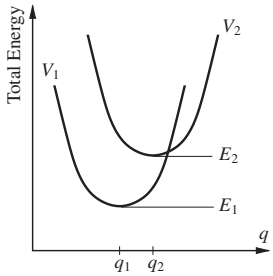

Fig. 49. The adiabatic potentials of the defect in its two states 1 and 2. The adiabatic potentials depend solely on the reaction coordinate q where the electrons are assumed to adjust immediately. The electronic configuration (neutral vs. charged) determines which potential has to be considered. Electron–phonon coupling leads to a shift of the minima of the parabolas q_i and a change in the oscillation frequency ω_i.

As mentioned above, the most important aspect about the model potentials (77) is that the two charge states have different equilibrium positions q_i. This is due to the aforementioned phenomenon of *electron–phonon coupling*, the essence of which is that the number and distribution of the electrons determines the vibrational properties of the system. This is easy to understand by considering that the occupancy of a bond determines its 'strength'. By changing the occupancy/strength, the atoms will move to a different equilibrium position. Similarly, the vibrations of the atoms determines the electron distributions. As a consequence, electron–phonon coupling makes the description of the defect much more complicated since in principle the wavefunctions of the electron and phonon system have to be determined in a coupled manner. According to the conventionally employed Born–Oppenheimer approximation, however, electronic and nuclear motion can be treated separately [96], as it is assumed that the light electrons can adjust quickly to changes in the positions of the sluggish atoms. While this approximation appears reasonable around the minimum, it is questionable particularly during a transition from one state to another, because then the electronic wave functions would have to change instantaneously for instance from a bound to a free state, which is impossible [47]. Nonetheless, in the simplest model, the contribution to the total energy due to electron–phonon coupling is assumed to be a linear function of the displacement. In such a model only the equilibrium positions are displaced while the two parabolas are otherwise identical ($\omega_1 = \omega_2$). In a more realistic model, electron–phonon coupling is assumed to be quadratic, which results in displaced parabolas with different vibrational frequencies ω_i. In the SRH model electron–phonon coupling is ignored. However, as we shall see, it essentially determines both the bias- as well as the temperature-dependence of the transition probabilities.

In order to quantum–mechanically describe the transition from one charge state to the other, one has to consider the occupancies of the eigenstates of the oscillator associated with the current state of the defect and calculate the overlap of the wavefunctions with the new state. The occupancies of the various states of this dominant mode are given by statistical physics while the energy levels have to be found from a solution of the Schrödinger equation [69,47]. The quantum–mechanical solution starts with the calculation of the eigenenergies for each harmonic oscillator, which are given by $\mathcal{E}_i = i\hbar\omega_1 + 1/2$ and $\mathcal{E}_j = j\hbar\omega_2 + 1/2$. For each pair of (i, j) the overlap of the corresponding wavefunctions is calculated to determine the transition probability according to Fermi's Golden Rule. However, as typical values of $\hbar\omega$ are around 20 meV and since for practical purposes we are mostly interested in the behavior of the defects above room temperature where $\hbar\omega$ is on the order of k_BT, a full quantum–mechanical solution is often not required and the defect behavior can be considered in a semiclassical approximation. The essence of the semiclassical approximation is that nuclear tunneling between the two defect states, which is only relevant at low temperatures anyway, is neglected.

Two kinds of transitions are important and have been discussed at lengths in the literature: first, radiative transitions, where energy is supplied via radiation (photons), and second, non-radiative transitions, where the energy is internally supplied via phonons.

6.1. Radiative transitions

Consider first radiative (or optical) transitions, which, according to the (classical) Franck–Condon, principle occur around the minima of the parabolas, see Fig. 50. The Franck–Condon principle states that during a transition the lattice coordinate q does not change. For example, lets assume that the defect is in state 1. Then, in order to have a transition to state 2, a photon has to supply the energy $\mathcal{E}_{12}^{\circ}$ to raise the energy of the system from E_1 to $E_1 + \mathcal{E}_{12}^{\circ}$. According to Fig. 50, this energy is larger than what one would expect from the purely electronic (SRH) picture, where the barrier is determined only by the difference in the energy levels, $E_{21} = E_2 - E_1$. This excess energy

$$\mathcal{E}_{R12} = \mathcal{E}_{12}^{\circ} - E_{21} = V_2(q_1) - V_2(q_2) \tag{78}$$

is known as the relaxation energy [95] and has to be dissipated via the emission of (many) phonons, hence the name *multiphonon* process.

The reverse process can occur as well: once the system is in its new equilibrium q_2, a transition to state 1 may occur, whereby the excess energy $\mathcal{E}_{21}^{\circ}$ is emitted via a photon. Quite remarkably, the energy of the emitted photon is smaller than that of the absorbed photon, a puzzling experimentally observable phenomenon which originally led to the development of multiphonon theory a 80 years ago, see e.g. Ref. [97]. Also, the relaxation energy $\mathcal{E}_{R12}$ only equals $\mathcal{E}_{R21}$ when $\omega_1 = \omega_2$.

The energy required to induce a radiative transition $(\mathcal{E}_{12}^{\circ})$ and the radiative energy released during the backward transition $(\mathcal{E}_{21}^{\circ})$ can be obtained from the binding energy, $E_B(q) = V_2(q) - V_1(q)$ for a given q. E_B can be easily calculated from (77) as

$$E_B(q) = E_{21} + \mathcal{E}_{R12} - q\frac{\sqrt{2M\omega_1^2\mathcal{E}_{R12}}}{R} + q^2\frac{M\omega_1^2}{2}\frac{1-R^2}{R^2}, \tag{79}$$

with $R = \omega_1/\omega_2$ and $\mathcal{E}_{R12} = S\hbar\omega_1 = M\omega_1^2 q_2^2/2$, where without loss of generality q_1 is set to zero. The parameter S is known as the Huang–Rhys factor and determines the number of phonons required to excite the harmonic oscillator in state 1 to reach the

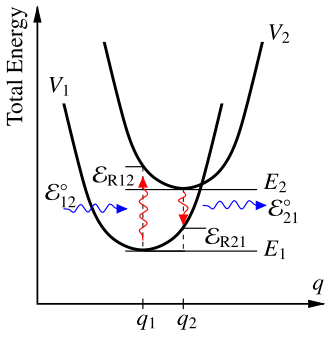

Fig. 50. A nonradiative (optical) transition between the states 1 and 2. Without electron–phonon coupling q_1 would equal q_2 and the absorbed energy $\mathcal{E}_{12}$ would equal the emitted energy $\mathcal{E}_{21}$. When electron–phonon coupling is not negligible, $\mathcal{E}_{12} \neq \mathcal{E}_{21}$. In general, the relaxation energies $\mathcal{E}_{R12}$ and $\mathcal{E}_{R21}$ are different, unless $\omega_1 = \omega_2$.

equilibrium position of state 2, q_2. From the binding energy we obtain the optical energies as

$$\mathcal{E}^{\circ}_{12} = E_B(0) = E_{21} + \mathcal{E}_{R12}, \tag{80}$$
$$\mathcal{E}^{\circ}_{21} = E_B(q_2) = E_{21} - \mathcal{E}_{R21}, \tag{81}$$

with $\mathcal{E}_{R21} = R^2\mathcal{E}_{R12}$. For $R = 1$, the relaxation energies are the same in both states, $\mathcal{E}_R = \mathcal{E}_{R21} = \mathcal{E}_{R12} = Sh\omega$.

Classically, at zero absolute temperature, the system will either be in state q_1 with energy E_1 or state q_2 with energy E_2. When the temperature is increased, higher energies $\mathcal{E}$ are occupied with probability

$$\frac{P(\mathcal{E})}{P(E_i)} = \frac{e^{-\beta\mathcal{E}}}{e^{-\beta E_i}} = e^{-\beta(\mathcal{E}-E_i)}. \tag{82}$$

The excess energy $\mathcal{E} - E_i$ is supplied by phonons. As a consequence, absorption and emission will not just occur at the minima q_1 and q_2, but in a region around them. The occurrence of these other transitions has an exponentially decreasing probability due to the Boltzmann factor (82). This leads to a thermal broadening of the absorption and emission lines with a maximum at q_1 and q_2, respectively.

6.2. Nonradiative transition barriers

Suppose now that no photons are available, as is usually the case when defects are studied during the regular operation of semiconductor devices. Without photons, direct transitions are not possible. Hence, we have to completely rely on phonons to increase the energy of the system sufficiently to allow a transition. As energy has to be conserved, a classical transition is only possible at the intersection point of the parabolas. There, the binding energy is zero, see Fig. 51. For instance, a transition from state 1 to state 2 requires the system to surmount the barrier $\mathcal{E}_{12}$. The probability that this state is occupied is given by the Boltzmann factor, (82), and we have for the capture coefficient

$$c_p(E) = v_{th}\sigma e^{-\beta\mathcal{E}_{12}}. \tag{83}$$

Using the equilibrium detailed balance relation (67) and Boltzmann statistics, we can write the two rates as

$$k_{12} = p v_{th}\sigma\langle\langle e^{-\beta\mathcal{E}_{12}}\rangle\rangle, \tag{84}$$
$$k_{21} = N_V v_{th}\sigma\langle\langle e^{-\beta\mathcal{E}_{12}} e^{\beta(E_V-E_1)}\rangle\rangle. \tag{85}$$

For this example defect, an electron is bound at the defect site in the neutral state 1. During the transition to state 2, this electron is transferred to the reservoir, usually the channel of the transistor. The potential energy of this electron in state 1 is given by E_1, while in state 2 it equals E_2. For simplicity, we assume in the following that $E_2 = E_V$, that is, the electron is moved back into the bottom of the silicon valence band. This assumption effectively removes the averaging over the slightly different barriers found at different values of E_2. Furthermore, since quite obviously $\mathcal{E}_{12} = \mathcal{E}_{21} + E_{21}$ for any shape of the parabolas, we have

$$k_{12} \approx p v_{th}\sigma e^{-\beta\mathcal{E}_{12}}, \tag{86}$$
$$k_{21} \approx N_V v_{th}\sigma e^{-\beta\mathcal{E}_{21}}. \tag{87}$$

As before in the radiative case, the barrier $\mathcal{E}_{12}$ can be calculated from the adiabatic potentials (77), which is a more-or-less straight forward exercise and will be done in the following.

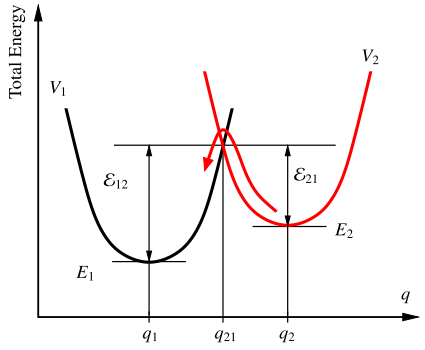

Fig. 51. When no photons are available, only nonradiative multiphonon transitions are possible. The necessary energy has to be supplied by phonons.

6.2.1. Linear electron–phonon coupling

As the calculation of $\mathcal{E}_{12}$ for the general case with $\omega_1 \neq \omega_2$ and $q_1 \neq q_2$ is a little awkward, we start with the simplest case where the vibrational frequencies are assumed to be the same in both states, that is, $\omega_1 = \omega_2$, or, equivalently, $R = 1$. Then, the quadratic term in q of $E_B(q)$ vanishes and only the term linear in q remains. This is known as *linear electron–phonon coupling* and is the most commonly discussed case, because it is the only case that allows for relatively compact solutions of the quantum–mechanical problem in the Born–Oppenheimer approximation [47]. Also, the classical solution is obtained in a nearly trivial manner. For $R = 1$ the two parabolas intersect at

$$q_{21} = \frac{\mathcal{E}_R + E_{21}}{\sqrt{2M\mathcal{E}_R\omega^2}}, \tag{88}$$

which then gives the well-known result

$$\mathcal{E}_{12} = \frac{(\mathcal{E}_R + E_{21})^2}{4\mathcal{E}_R}, \tag{89}$$

for the barrier separating state 2 from state 1. Analogously, the barrier for the backward transition is obtained as

$$\mathcal{E}_{21} = \frac{(\mathcal{E}_R - E_{21})^2}{4\mathcal{E}_R}, \tag{90}$$

where $\mathcal{E}_{21} = \mathcal{E}_{12} - E_{21}$ holds.

We have seen previously that E_{21} depends linearly on the electric field F via (74), see Fig. 52. Thus, at a first glance, (89) suggests that $\mathcal{E}_{12}$ depends *quadratically* on F. This, however, is not the case because electron–phonon coupling is usually strong, that is, $\mathcal{E}_R \gg E_{21}$. In that case the squares Eqs. (89) and (90) can be expanded [98] to yield

$$\mathcal{E}_{12} \approx \frac{1}{4}\mathcal{E}_R + \frac{1}{2}E_{21}, \tag{91}$$
$$\mathcal{E}_{21} \approx \frac{1}{4}\mathcal{E}_R - \frac{1}{2}E_{21}. \tag{92}$$

This shows that the apparent activation energy is dominated by the relaxation energy and $E_A \approx \mathcal{E}_R/4$.

In contrast, if the electron–phonon coupling is weak ($\mathcal{E}_R \ll E_{21}$), expansion of (89) gives

$$\mathcal{E}_{12} \approx \frac{E_{21}^2}{4\mathcal{E}_R} + \frac{1}{2}E_{21}. \tag{93}$$

The first term is usually dominant and one obtains a quadratic dependence on F. So the apparent quadratic F-dependence in (89) is only visible in the weak electron–phonon coupling limit $\mathcal{E}_R \ll E_{21}$.

Comparison of the forward and backward barriers in the strong coupling case, (91) and (92), already highlights a significant difference between the SRH and the NMP model: in the SRH model, the barrier E_{21} and thus the F-dependence enters either the capture or the emission time constant. So, when then capture time constant depends exponentially on F, the emission time constant will be bias independent and vice versa. Quite differently, in the NMP model, the bias dependence is equally shared with different signs

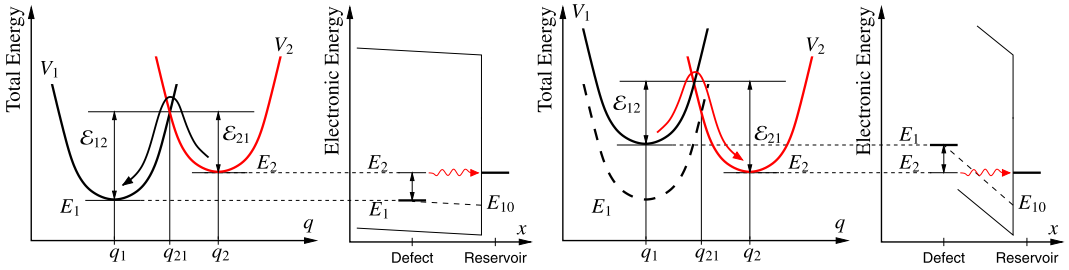

Fig. 52. The field-dependence of the NMP transition rates is a consequence of the electrostatic shift of the defect level.

between the time constants. For instance, if the capture time constant decreases exponentially as $\exp(-x\,F/2V_T)$, the emission time constant will increase exponentially as $\exp(x\,F/2V_T)$.

The barriers for the general case together with the two limiting cases are summarized in Table 1. The qualitative differences between the strong and weak coupling regimes are shown in Fig. 53.

6.2.2. Quadratic electron–phonon coupling

As long as we remain within the semiclassical approximation, the situation is only slightly more complicated for the general case where both the linear and quadratic term are considered. Contrary to the case $R = 1$, where only one intersection of the parabolas exists, for $\omega_i \neq \omega_j$ the two parabolas have now either two crossings or none. The case non-intersecting parabolas has been discussed previously to explain unexpectedly small capture cross sections [47]. Naturally, in such a case the semiclassical approximation can no longer be used for the calculation of the transition probabilities and a quantum–mechanical solution has to be employed [47,99,100]. On the other hand, for the case that two crossings exists, we will assume that the smaller barrier dominates the transitions, which reads

$$\mathcal{E}_{12} = \frac{\mathcal{E}_{R21}}{(R^2-1)^2}\left(1 - R\sqrt{1 + (R^2-1)\mathcal{E}_{21}/\mathcal{E}_{R21}}\right)^2. \tag{94}$$

Eq. (94) is somewhat awkward to use, particularly due to the removable singularity at $R = 1$, which corresponds to a negligible quadratic coupling term. For strong electron–phonon coupling, $E_{21} \ll \mathcal{E}_{R21}$, we can expand (94) up to second-order in E_{21} to find

$$\mathcal{E}_{12} \approx \frac{\mathcal{E}_{R21}}{(1+R)^2} + \frac{R}{1+R}E_{21} + \frac{R}{4\mathcal{E}_{R21}}E_{21}^2. \tag{95}$$

This expansion has the advantage that it contains the correct limit (89) when $R = 1$ and thus lends itself nicely to analytic treatments. From $\mathcal{E}_{21} = \mathcal{E}_{12} - E_{21}$ we obtain the barrier for the reverse reaction as

$$\mathcal{E}_{21} \approx \frac{\mathcal{E}_{R21}}{(1+R)^2} - \frac{1}{1+R}E_{21} + \frac{R}{4\mathcal{E}_{R21}}E_{21}^2. \tag{96}$$

An important conclusion that can be drawn from (95) is that the oscillation frequency mismatch R also enters the apparent activation energy, which is roughly $\mathcal{E}_{R21}/(1+R)^2$. Also, the field dependence is now no longer symmetric. For example, for $R > 1$, $\mathcal{E}_{12}$ now depends stronger on F than $\mathcal{E}_{21}$ and vice versa.

6.2.3. Problems of the model

While the NMP model captures the 'essence', like a strong bias- and temperature-dependence, important details seen experimentally are missing. First, the bias dependence of τ_c and τ_e are symmetric, at least in the linear electron–phonon coupling mode. Furthermore, the predicted bias-dependence of τ_c is nearly linear. As shown in Fig. 28, however, the bias-dependence of τ_c has some curvature on a logarithmic scale. Also, as shown in Fig. 29, τ_e is normally bias-independent above V_{th} but may drop abruptly below V_{th}. So even if quadratic electron–phonon coupling is taken into account, which allows to introduce some asymmetry between τ_c and τ_e, this cannot be driven far enough to make τ_e nearly bias-independent within a meaning full range of oscillation frequencies. Also, the rapid drop of τ_e in some defects below V_{th} remains puzzling. Finally, although the situation is considerably improved compared to the SRH model, no full decorrelation between τ_c and τ_e is possible: A typical CET map produced by the NMP model with a Gaussian distribution of E_1 and a uniform distribution of x is given in Fig. 54. While τ_c and τ_e are no longer correlated, for a given τ_c the predicted τ_e remains within a relatively narrow band, given by the active area.

6.2.4. Kirton and Uren model

From a historical perspective, the model employed in the pioneering work by Kirton and Uren [50] requires to be mentioned, as it appears to be still widely used. Already in those days it was recognized that the SRH model is unable to explain the experimental data and that a nonradiative multiphonon process could be responsible. In order to account for this, Kirton and Uren introduced a Boltzmann factor into the SRH rates to account for structural relaxation. While this introduces the missing temperature-dependence into the model, it is not rigorously correct, since the Boltzmann factors of the NMP model summarized in Table 1 are also strongly

Table 1
Forward and backward barriers of the NMP model for $\omega_1 = \omega_2$. Even for weak-coupling the NMP model behaves quite differently compared to the SRH model.

Model	$\mathcal{E}_{12}$	$\mathcal{E}_{21}$
NMP ($R = 1$)	$\frac{(\mathcal{E}_R + E_{21})^2}{4\mathcal{E}_R}$	$\frac{(\mathcal{E}_R - E_{21})^2}{4\mathcal{E}_R}$
NMP ($R = 1$) weak coupling ($\mathcal{E}_R \ll E_{21}$)	$E_{21}\left(\frac{E_{21}}{4\mathcal{E}_R} + \frac{1}{2}\right)$	$E_{21}\left(\frac{E_{21}}{4\mathcal{E}_R} - \frac{1}{2}\right)$
NMP ($R = 1$) strong coupling ($\mathcal{E}_R \gg E_{21}$)	$\frac{\mathcal{E}_R}{4} + \frac{E_{21}}{2}$	$\frac{\mathcal{E}_R}{4} - \frac{E_{21}}{2}$
NMP ($R \neq 1$) strong coupling ($\mathcal{E}_{R21} \gg E_{21}$)	$\frac{\mathcal{E}_{R21}}{(1+R)^2} + \frac{R}{1+R}E_{21}$	$\frac{\mathcal{E}_{R21}}{(1+R)^2} - \frac{1}{1+R}E_{21}$
SRH	$\max(E_{21}, 0)$	$\max(-E_{21}, 0)$

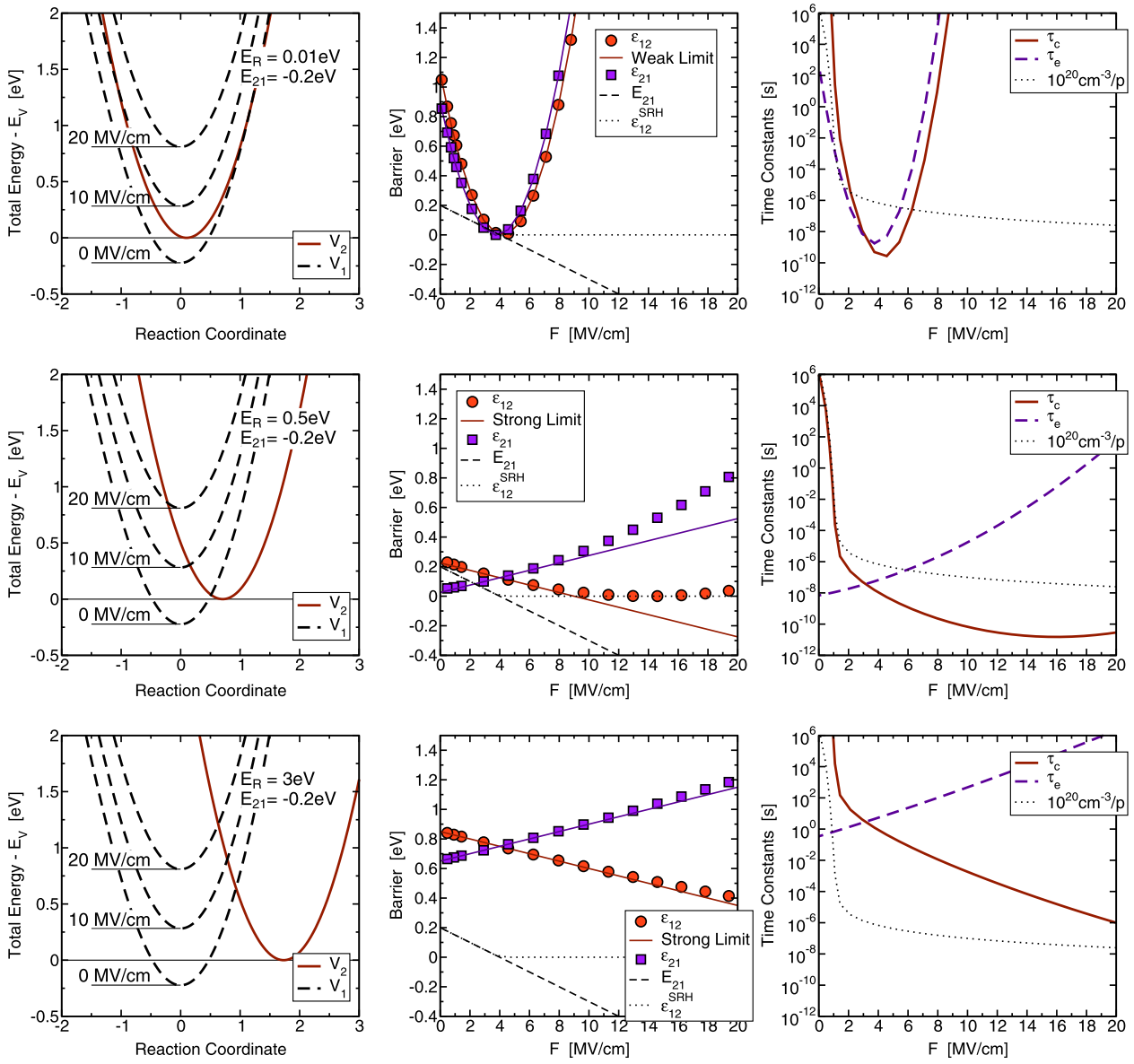

Fig. 53. Comparison of the weak (top), intermediate (middle), and strong (bottom) electron–phonon coupling regimes. The left panels show the adiabatic potentials for three difference values of the Huang–Rhys factor S. The only impact of S is that with increasing S the potential of the positive state V_2 is rigidly shifted to larger values of the reaction coordinate. The middle panels show the forward and backward barriers as a function of the electric field. Also shown are the energies E_{12} and the barrier used in the SRH model, which behaves quite differently. Finally, the rightmost panels show the resulting capture and emission time constants. Note that the strong sensitivity below $F = 2$ MV/cm is due to the hole concentration (dotted lines) rather than the barriers.

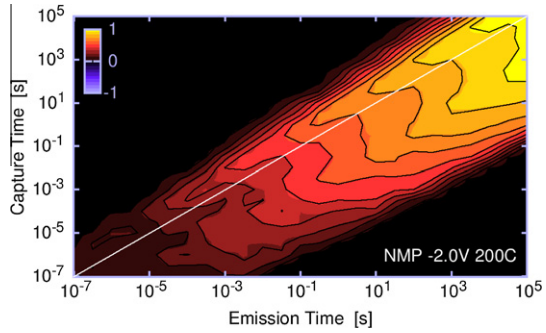

Fig. 54. The CET map obtained from the NMP model with a Gaussian distribution of E_1 and a uniform distribution of x into the oxide. The width of the distribution is determined by the active area and therefore narrower than observed experimentally.

bias-dependent. Furthermore, their approach implicitly assumed that the energy levels of the defect are inside the silicon bandgap.

7. Multi-state defect model

In order to correct for the limitations of the simple NMP model we recall the existence of metastable defect states, which show up as either anomalous RTN, temporary RTN, and temporarily disappearing defects. As it turns out, consideration of these metastable states naturally solves many issues identified with the simple NMP model.

We construct our multi-state defect model based on reported properties of the E' center. It has been suggested that a fraction of the E' centers created following irradiation tests can be repeatedly charged and discharged. The corresponding energy levels lie within the silicon bandgap [101]. The idea behind this cyclability

is that once the hole is emitted (that is, an electron is captured), the defect does not fully relax but remains in a metastable state which can easily lose an electron again. The fact that it is the E' center that can act as a switching trap has been suggested by Lelis et al. [102] based on electrical measurements. This was later confirmed by ESR studies [5] and theoretical calculations [103]. In Ref. [93] it has been suggested that in order to create a stable E' configuration from an oxygen vacancy, the doubly positive configuration could be important. Diverging opinions have been expressed in Refs. [7,8], where the switching behavior was put down to hydrogen which can change its charge state in the amorphous network.

No matter what the microscopic defect configuration is, the model can be formulated in an 'agnostic' fashion by considering four states as schematically shown in Fig. 55. Starting from the equilibrium state 1, hole capture transforms the defect into the metastable state 2′, where the distance between the silicon atoms is only slightly larger than in the neutral state 1. Depending on the defect configuration, state 2′ may be only metastable, that is, can relax into a more stable form by an increase of the distance between the silicon atoms. Thereby one silicon atom moves through the plane spanned by its three oxygen neighbors and forms a bond with an oxygen atom in its back. Strictly speaking, this is only a likely scenario in crystalline SiO_2, as in an amorphous network no suitable oxygen atom may be available. As such, although E' centers have been studied in great detail for many decades by now, the suitability of the E' center in the present model is anything but certain. While first principles calculations show that E' centers have the required metastable states with most barriers having suitable values [104,90]. In amorphous SiO_2 these barriers show a wide distribution. Unfortunately, though, the energy level of the E' center in crystalline SiO_2 is only about 1 eV above the SiO_2 valence band, which is too low to result in reasonable capture rates. On the other hand, the thermodynamic energy level of the hydrogen bridge was found to be in the middle of the Si bandgap, meaning that most defects would already be positively charged prior to stress. Finally, while the positively charged, puckered state (state 2) was found to be stable, the barrier from the electrically neutralized puckered state (state 1′) to the equilibrium state (state 1) was found to be very small. This would be inconsistent with defects that can be repeatably charged and discharged. Apart from the reasons listed above, a number of other explanations for this

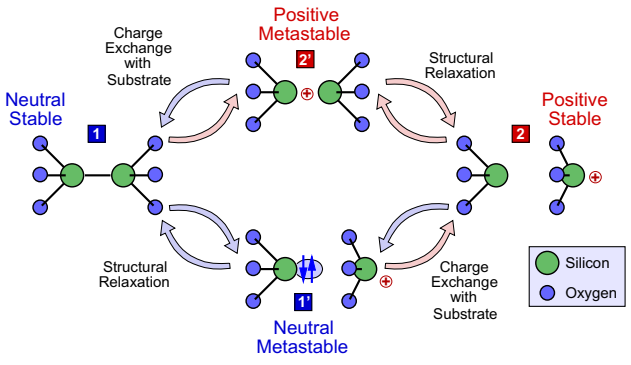

Fig. 55. The four states of a switching oxide trap, using the E' center as an example. Initially, a neutral precursor exists (state 1). Upon hole capture, the Si–Si bond breaks and a positively charged E' center is created (state 2′). Depending on the defect configuration, state 2′ may transform into the puckered configuration (2) when the silicon atom moves through the plane spanned by its three neighboring oxygen atoms. Hole emission (electron capture) neutralizes the E' center (state 1′). Being in state 1′, two options exist: a hole can be captured again, causing a transition back to state 2, or the structure can relax back to its equilibrium configuration (state 1).

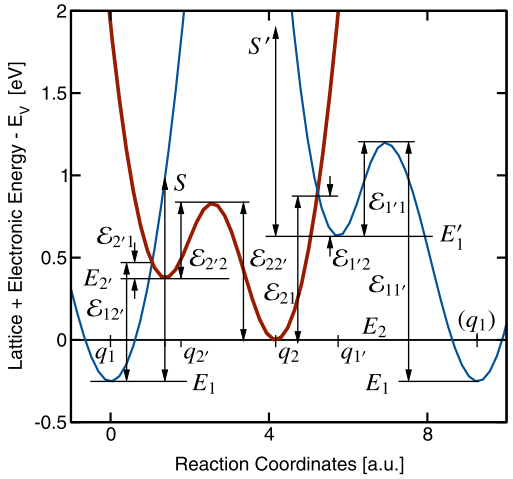

Fig. 56. Definition of the symbols used in the model. Note that the reaction coordinate describing the transition $1 \leftrightarrow 2'$ is different from the one describing $2 \leftrightarrow 1'$, which is why the potential describing states 1 and 1′ is plotted twice, once to the left and once to the 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 $q_{1'}$ determine $2 \leftrightarrow 1'$.

mismatch exist and a considerable amount of work remains to be done in order to clarify these issues.

In order to setup the transition rates for the model shown in Fig. 55, we use the adiabatic potentials shown in Fig. 56. Each charge state is now represented by a double well, with the equilibrium configuration being the minimum (1 and 2) and the metastable states as the energetically higher value. Transitions including charge transfer are modeled analogously to Eqs. (86) and (87) using nonradiative multiphonon theory. Transitions without change of the charge state ($1 \leftrightarrow 1'$ and $2 \leftrightarrow 2'$) are assumed to follow a simple thermal excitation over a barrier. Thus, the charge transfer rates are

$$k_{12'} = p\sigma v_{\mathrm{th}} e^{-\beta\mathcal{E}_{12'}}, \qquad k_{2'1} = N_{\mathrm{V}}\sigma v_{\mathrm{th}} e^{-\beta\mathcal{E}_{2'1}}, \tag{97}$$

$$k_{1'2} = p\sigma v_{\mathrm{th}} e^{-\beta\mathcal{E}_{1'2}}, \qquad k_{21'} = N_{\mathrm{V}}\sigma v_{\mathrm{th}} e^{-\beta\mathcal{E}_{21'}}. \tag{98}$$

For simplicity we assume all capture cross sections to be equal. The transition between states 1 and 1′ as well as 2′ and 2 are assumed 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, $\mathcal{E}_{22'} = \mathcal{E}_{2'2} + E_{2'} - E_2$ and $\mathcal{E}_{11'} = \mathcal{E}_{1'1} + E'_1 - E_1$ and we use $k_{mn} = \nu_m \exp(-\beta\mathcal{E}_{mn})$ where $\nu_m \sim 10^{13}\ \mathrm{s}^{-1}$.

7.1. Approximate solutions

As discussed in Section 2.7, the solution of the master Eq. (31) is in principle straight forward to obtain for this four-state defect. However, it does not provide significant insight into the behavior of the defect due to its complexity. 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. 57. 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 $|V_{\mathrm{G}}|$. Thus, we will make use of the 'effective two-state' defect model developed in Section 2.7. We again only consider the

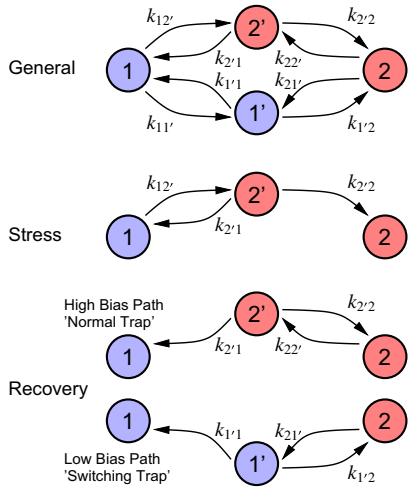

Fig. 57. State transition rate diagram for the general case (top), stress case (middle), and the two important pathways for recovery (bottom). During stress and recovery the Markov chain is described by a Birth–Death process since we are only interested in the first passage time from the initial to the final state, the expectation value of which gives the time constant.

dominant charge exchange with the valence band in the substrate. Also, the second-order term in E_{12} in the expansions Eqs. (95) and (96) will be neglected. Finally, Boltzmann statistics will be assumed again 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. 57. 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. 57). 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

$$1/\bar{\tau}_c = 1/\bar{\tau}_c^{2'} + 1/\bar{\tau}_c^{1'}. \tag{99}$$

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

$$1/\bar{\tau}_e = 1/\bar{\tau}_e^{2'} + 1/\bar{\tau}_e^{1'}. \tag{100}$$

The individual contributions to the time constants are

$$\bar{\tau}_c^{2'} = \bar{\tau}_{c;\min}^{2'}\left(1 + \frac{N_1}{p}\exp\left(-\frac{xF}{V_T}\right)\right) + \tau_0\frac{N_2}{p}\exp\left(-\frac{xR}{1+R}\frac{F}{V_T}\right), \tag{101}$$

$$\bar{\tau}_c^{1'} = \bar{\tau}_{c;\min}^{1'} + \tau_0\frac{N_3}{p}\exp\left(-\frac{xR'}{1+R'}\frac{F}{V_T}\right), \tag{102}$$

$$\bar{\tau}_e^{2'} = \bar{\tau}_{e;\min}^{2'} + \tau_{2'}\exp\left(\frac{x}{1+R}\frac{F}{V_T}\right), \tag{103}$$

$$\bar{\tau}_e^{1'} = \bar{\tau}_{e;\min}^{1'}\left(1 + e^{\beta E_{1'F}}\right) + \tau_{1'}\exp\left(\frac{x}{1+R'}\frac{F}{V_T}\right), \tag{104}$$

with $\tau_0^{-1} = N_V \nu_{th}\sigma_0 e^{-x/x_0}$ and the temperature-dependent but field-independent auxiliary quantities

$$N_1 = N_V\exp\left(\beta(E_{2'2} - \Delta E_1)\right),$$

$$N_2 = N_V\exp\left(\beta\left(\frac{\mathcal{E}_R}{(1+R)^2} - \frac{R(\Delta E_1 - E_{2'2})}{1+R}\right)\right),$$

$$N_3 = N_V\exp\left(\beta\left(\frac{\mathcal{E}'_R}{(1+R')^2} - \frac{R'\Delta E_{1'}}{1+R'}\right)\right)\left(1+\exp\left(\beta(\Delta E_{1'} - \Delta E_1)\right)\right),$$

$$\tau_{2'} = \tau_0\exp\left(\beta\left(\frac{\mathcal{E}_R}{(1+R)^2} + \frac{\Delta E_1 - E_{2'2}}{1+R}\right)\right)\left(1+\exp\left(\beta E_{2'2}\right)\right),$$

$$\tau_{1'} = \tau_0\exp\left(\beta\left(\frac{\mathcal{E}'_R}{(1+R')^2} - \frac{\Delta E_{1'}}{1+R'}\right)\right),$$

with $\Delta E_1 = E_{10} - E_{V0}$, $\Delta E_{1'} = E_{1'0} - E_{V0}$, $R = \omega_1/\omega_{2'}$, and $R' = \omega_{1'}/\omega_2$. As each time constant contains a contribution from a purely thermal transition, the minimum value is bounded by

$$\bar{\tau}_{c;\min}^{2'} = 1/k_{2'2}, \bar{\tau}_{c;\min}^{1'} = 1/k_{11'}, \bar{\tau}_{e;\min}^{2'} = 1/k_{22'}, \quad \text{and } \bar{\tau}_{e;\min}^{1'} = 1/k_{1'1}.$$

Although the simple Eqs. (99) and (100) are not rigorously correct, since they simply superposition two independent three-state defects to approximate the behavior of the four-state defect, they give a very good approximation. In order to demonstrate this, an evaluation of the analytic capture and emission time models against a Monte Carlo simulation of the full model is given in Fig. 58 for a switching trap, where around $V_G = V_{th}$ the dominant pathway changes from via 2′ to 1′.

As the analytic solution (101)–(104) for the effective capture and emission time constants is still rather formidable, it is worthwhile to explore two limiting cases in the following. Which case becomes relevant depends on the adiabatic defect potentials, which are expected to be different for each defect.

7.1.1. 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 $\Delta E_{1'}$ is too large to give a significant occupancy of state 1′. In this case, the time constants are

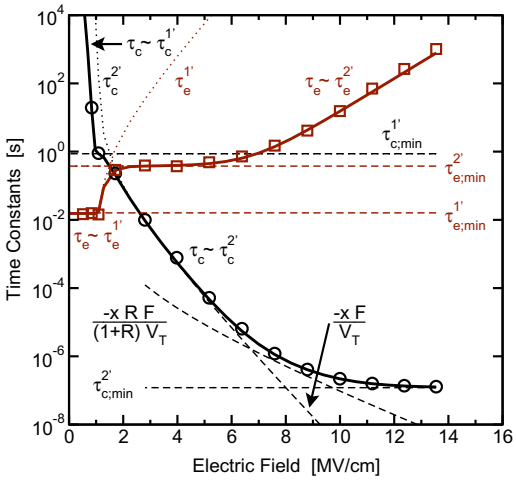

Fig. 58. Bias dependence of the effective capture and emission times according to the model. The symbols are obtained from a Monte Carlo simulation of the stochastic process, while the dotted lines give the individual contributions to the effective time constants (solid lines).

$$\bar{\tau}_c = \bar{\tau}^{2'}_{c;\min}\left(1+\frac{N_1}{p}\exp\left(-\frac{xF}{V_T}\right)\right)+\tau_0\frac{N_2}{p}\exp\left(-\frac{xR}{1+R}\frac{F}{V_T}\right), \quad (105)$$

$$\bar{\tau}_e = \bar{\tau}^{2'}_{e;\min}+\tau_{2'}\exp\left(\frac{x}{1+R}\frac{F}{V_T}\right). \quad (106)$$

Both time constants consist now of two terms, where the first one denotes the impact of the relaxation barrier $\mathcal{E}_{2'2}$. For capture, the field dependence of the two terms is different, resulting in a non-linear overall exponential field dependence and eventual saturation at $\bar{\tau}_c = \bar{\tau}^{2'}_{c;\min}$ for high fields. For example, for $R = 1$, the argument of the exponential field term is initially $-x\,F/V_T$ and gradually reduces to $-x\,F/2V_T$ 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.

For emission, the term due to the relaxation barrier is bias-independent and dominates for small fields. This is similar to the SRH model where irrespective of the exponential F dependence only a weak field dependence is obtained, as F depends weakly on V_G below V_{th}.

7.1.2. 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, the transition 2 → 1′ can occur and the defect may even switch back and forth between states 1′ and 2, see Fig. 57 (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. 58. 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 → 1′ → 2 is theoretically possible, we have so far not observed a defect compatible with such a configuration.

Introducing $f_n = (1+k_{1'2}/k_{2'1})^{-1} = (1+\exp(\beta E_{1'F}))^{-1}$, which is the probability that the trap level E'_1 is occupied by an electron, that is, neutral, we can express the emission time constant for $\bar{\tau}^{1'}_e \lesssim \bar{\tau}^{2'}_e$ as

$$\bar{\tau}_e = \frac{\bar{\tau}^{1'}_{e;\min}}{f_n}+\tau_{1'}\exp\left(\frac{x}{1+R'}\frac{F}{V_T}\right). \quad (107)$$

This is a remarkable result: when $E_F > E'_1$, the defect is neutral ($f_n = 1$) and the emission time is given by the bias independent value $\bar{\tau}^{1'}_{e;\min}$. As soon as the defect level E'_1 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. 29. For large $|V_G|$, on the other hand, $\bar{\tau}^{1'}_e$ will become larger than $\bar{\tau}^{2'}_e$ and the pathway 2 → 2′ → 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 7.2.

7.2. Qualitative model behavior

A stochastic simulation of a defect configuration leading to tRTN is shown in Fig. 59. 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

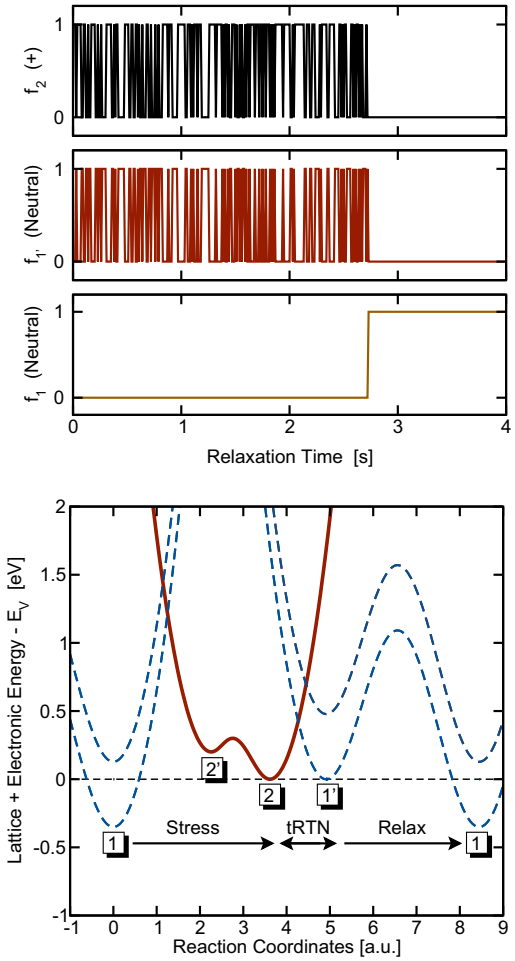

Fig. 59. Stochastic simulation (top) of a defect configuration (bottom) that leads to tRTN. During stress, state 2 is occupied, during the initial recovery phase the defect switches between states 2 and 1′, while eventually the defect anneals. Such a behavior can only be observed in a very narrow window of V_G where the minima of states 2 and 1′ are close to each other.

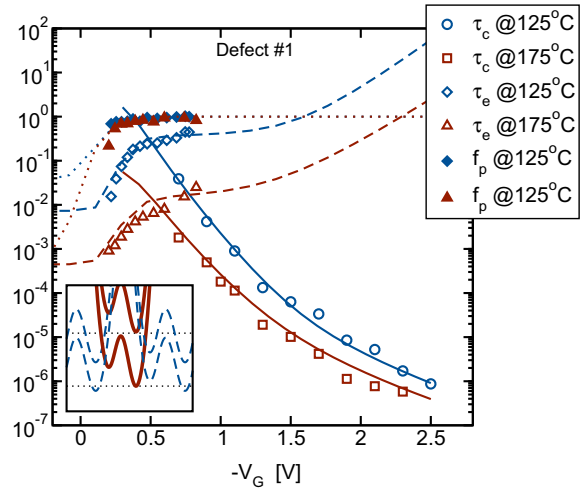

Fig. 60. Simulated capture and emission time constants for defect #1 compared to the experimental values. The CC diagram shown in the inset is similar to that of the tRTN case shown in Fig. 59 with the difference that the barrier between states 2 and 1′ is rather small. The experimental occupation probability f_p is given by the filled symbols.

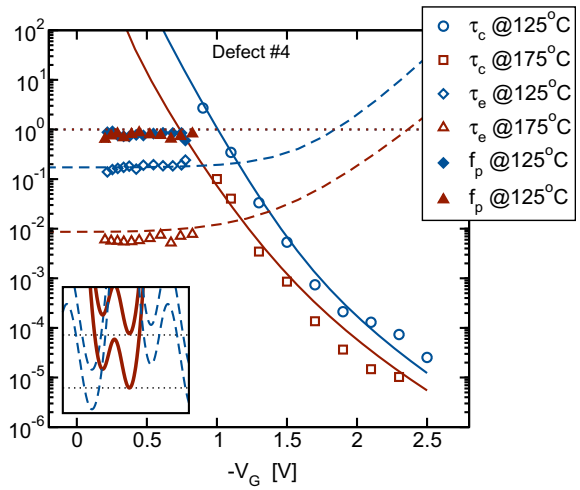

Fig. 61. Similar to Fig. 60, but now for defect #4. In contrast to defect #1, E_1' is too high to be significantly populated and no switching trap behavior can be observed. This can be clearly seen in the CC diagram shown in the inset. As a consequence, the emission time constant is insensitive to the gate bias.

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.

7.3. Quantitative model behavior

The simulated capture and emission time constants calibrated to the experimental data available for defect #1 are shown in Fig. 60. As can be seen from the CC diagram shown in the inset, defect #1 is similar to the schematic tRTN case shown in Fig. 59 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. 61 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.

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.

8. Conclusions

Particularly in the reliability community, oxide charge trapping is still dominantly interpreted using extended Shockley–Read–Hall-like models. A detailed analysis reveals, however, that these models are often at odds with experimental data, which are better described using nonradiative multiphonon models. Strictly speaking, a quantum–mechanical solution of this problem is required, which is rather involved for realistic defect potentials. Fortunately, the classical approximation of nonradiative multiphonon processes is accurate enough for practical purposes above room temperature. As such, very intuitive and simple expressions for the capture and emission rates can be obtained which are not significantly more complicated than those of the popular SRH model.

This review tries to provide a summary of the fundamentals required to go beyond the SRH picture in order to understand more realistic models. In addition to multiphonon processes, one of the fundamental ingredients of such a model appears to be the fact that oxide defects can have more than two states. These additional states cause all sorts of interesting defect behavior, including anomalous and temporary RTN, disappearing defects, and a more complicated bias-dependence of the capture and emission times. Most intriguingly, these defect models provide a link between RTN and the bias temperature instabilities.

Acknowledgments

This work would not have been possible without the invaluable support of B. Kaczer (imec), H. Reisinger (Infineon Munich), Th. Aichinger/M.Nelhiebel (KAI/Infineon), and R. Minixhofer/H. Enichlmair (austriamicrosystems). Furthermore, inspiring discussions with W. Goes, Ph. Hehenberger, P.-J. Wagner, F. Schanovsky, M. Bina, J. Franco, M. Toledano-Luque, P. Lenahan, J. Campbell, J. Ryan, G. Pobegen, D. Gillespie, C. Schlünder, V. Huard, N. Mielke, G. Bersuker, B. Knowlton, R. Southwick, A.L. Shluger, A. Krishnan, S. Zafar, S. Mahapatra, E. Islam, and A. Alam are gratefully acknowledged.

The research leading to these results has received funding from the European Community's Seventh Framework Programme under Grant Agreement No. 216436 (ATHENIS) and from the ENIAC Project No. 820379 (MODERN).

References

[1] Deal B. Standardized terminology for oxide charges associated with thermally oxidized silicon. IEEE Trans Electr Dev 1980;27(3):606–8.
[2] Fleetwood D. "Border Traps" in MOS devices. IEEE Trans Nucl Sci 1992;39(2):269–71.
[3] Grasser T, Reisinger H, Wagner P-J, Goes W, Schanovsky F, Kaczer B. The time dependent defect spectroscopy (TDDS) technique for the bias temperature instability. In: Proc IRPS; 2010. p. 16–25.
[4] Grasser T, Reisinger H, Wagner P-J, Kaczer B. The time dependent defect spectroscopy for the characterization of border traps in metal–oxide–semiconductor transistors. Phys Rev B 2010;82(24):245318.
[5] Conley Jr J, Lenahan P, Lelis A, Oldham T. Electron spin resonance evidence for the structure of a switching oxide trap: long term structural change at silicon dangling bond sites in SiO_2. Appl Phys Lett 1995;67(15):2179–81.
[6] Lenahan P, Conley Jr J. What can electron paramagnetic resonance tell us about the Si/SiO_2 system? J Vac Sci Technol B 1998;16(4):2134–53.
[7] de Nijs J, Druijf K, Afanas'ev V, van der Drift E, Balk P. Hydrogen induced donor-type Si/SiO_2 interface states. Appl Phys Lett 1994;65(19):2428–30.
[8] Afanas'ev V, Stesmans A. Proton nature of radiation-induced positive charge in SiO_2 layers on Si. Eur Phys Lett 2001;53(2):233–9.
[9] Fleetwood D, Xiong H, Lu Z-Y, Nicklaw C, Felix J, Schrimpf R, et al. Unified model of hole trapping, $1/f$ noise, and thermally stimulated current in MOS devices. IEEE Trans Electr Dev 2002;49(6):2674–83.
[10] Jeppson K, Svensson C. Negative bias stress of MOS devices at high electric fields and degradation of MNOS devices. J Appl Phys 1977;48(5):2004–14.
[11] Ogawa S, Shiono N. Generalized diffusion–reaction model for the low-field charge build up instability at the Si/SiO_2 interface. Phys Rev B 1995;51(7):4218–30.
[12] Kaczer B, Grasser T, Martin-Martinez J, Simoen E, Aoulaiche M, Roussel P, et al. NBTI from the perspective of defect states with widely distributed time scales. In: Proc IRPS; 2009. p. 55–60.
[13] Grasser T, Kaczer B, Goes W, Aichinger T, Hehenberger P, Nelhiebel M. A two-stage model for negative bias temperature instability. In: Proc IRPS; 2009. p. 33–44.
[14] Grasser T, Reisinger H, Goes W, Aichinger T, Hehenberger P, Wagner P, et al. Switching oxide traps as the missing link between negative bias temperature instability and random telegraph noise. In: Proc IEDM; 2009. p. 729–32.
[15] Neugroschel A, Bersuker G, Choi R, Cochrane C, Lenahan P, Heh D, et al. An accurate lifetime analysis methodology incorporating governing NBTI mechanisms in high-k/SiO_2 gate stacks. In: Proc IEDM; 2006. p. 1–4.
[16] Young C, Zhao Y, Heh D, Choi R, Lee B, Bersuker G. Pulsed I_d–V_g methodology and its application to electron-trapping characterization and defect density profiling. IEEE Trans Electr Dev 2008;56(6):1322–9.
[17] Zhao K, Stathis J, Linder B, Cartier E, Kerber A. PBTI under dynamic stress: from a single defect point of view. In: Proc IRPS; 2011. p. 372–80.
[18] Alam M. A critical examination of the mechanics of dynamic NBTI for pMOSFETs. In: Proc IEDM; 2003. p. 345–8.
[19] Islam A, Kufluoglu H, Varghese D, Mahapatra S, Alam M. Recent issues in negative-bias temperature instability: initial degradation, field dependence of interface trap generation, hole trapping effects, and relaxation. IEEE Trans Electr Dev 2007;54(9):2143–54.

[20] Grasser T, Goes W, Kaczer B. Dispersive transport and negative bias temperature instability: boundary conditions, initial conditions, and transport models. IEEE Trans Dev Mater Rel 2008;8(1):79–97.
[21] Huard V, Denais M, Parthasarathy C. NBTI degradation: from physical mechanisms to modelling. Microelectr Reliab 2006;46(1):1–23.
[22] Reisinger H, Blank O, Heinrigs W, Mühlhoff A, Gustin W, Schlünder C. Analysis of NBTI degradation- and recovery-behavior based on ultra fast V_{th}-measurements. In: Proc IRPS; 2006. p. 448–53.
[23] Grasser T, Goes W, Sverdlov V, Kaczer B. The universality of NBTI relaxation and its implications for modeling and characterization. In: Proc IRPS; 2007. p. 268–80.
[24] Grasser T, Goes W, Kaczer B. Towards engineering modeling of negative bias temperature instability. In: Fleetwood D, Schrimpf R, Pantelides S, editors. Defects in microelectronic materials and devices. Taylor and Francis/CRC Press; 2008. p. 399–436.
[25] Grasser T, Kaczer B, Goes W. An energy-level perspective of bias temperature instability. In: Proc IRPS; 2008. p. 28–38.
[26] Aichinger T, Nelhiebel M, Grasser T. Unambiguous identification of the NBTI recovery mechanism using ultra-fast temperature changes. In: Proc IRPS; 2009. p. 2–7.
[27] Teo Z, Ang D, See K. Can the reaction–diffusion model explain generation and recovery of interface states contributing to NBTI? In: Proc IEDM; 2009. p. 737–40.
[28] Huard V. Two independent components modeling for negative bias temperature instability. In: Proc IRPS; 2010. p. 33–42.
[29] Ang D, Teo Z, Ho T, Ng C. Reassessing the mechanisms of negative-bias temperature instability by repetitive stress/relaxation experiments. IEEE Trans Dev Mater Rel 2011;11(1):19–34.
[30] Grasser T, Kaczer B, Goes W, Reisinger H, Aichinger T, Hehenberger P, et al. The paradigm shift in understanding the bias temperature instability: from reaction–diffusion to switching oxide traps. IEEE Trans Electr Dev, in press.
[31] Grasser T, Kaczer B, Aichinger T, Goes W, Nelhiebel M. Defect creation stimulated by thermally activated hole trapping as the driving force behind negative bias temperature instability in SiO_2, SiON, and high-k gate stacks. In: IIRW final rep.; 2008. p. 91–5.
[32] Kaczer B, Grasser T, Roussel P, Martin-Martinez J, O'Connor R, O'Sullivan B, et al. Ubiquitous relaxation in BTI stressing-new evaluation and insights. In: Proc IRPS; 2008. p. 20–7.
[33] Grasser T, Kaczer B, Hehenberger P, Goes W, O'Connor R, Reisinger H, et al. Simultaneous extraction of recoverable and permanent components contributing to bias-temperature instability. In: Proc IEDM; 2007. p. 801–4.
[34] Grasser T, Aichinger T, Pobegen G, Reisinger H, Wagner P-J, Franco J, et al. The 'permanent' component of NBTI: composition and annealing. In: Proc IRPS; 2011. p. 605–13.
[35] Rangan S, Mielke N, Yeh E. Universal recovery behavior of negative bias temperature instability. In: Proc IEDM; 2003. p. 341–4.
[36] Haggag A, Anderson G, Parihar S, Burnett D, Abeln G, Higman J, et al. Understanding SRAM high-temperature-operating-life NBTI: statistics and permanent vs recoverable damage. In: Proc IRPS; 2007. p. 452–6.
[37] Aichinger T, Nelhiebel M, Grasser T. A combined study of p- and n-channel MOS devices to investigate the energetic distribution of oxide traps after NBTI. IEEE Trans Electr Dev 2009;56(12):3018–26.
[38] Aichinger T, Puchner S, Nelhiebel M, Grasser T, Hutter H. Impact of hydrogen on recoverable and permanent damage following negative bias temperature stress. In: Proc IRPS; 2010. p. 1063–8.
[39] Aichinger T, Nelhiebel M, Einspieler S, Grasser T. Observing two stage recovery of gate oxide damage created under negative bias temperature stress. J Appl Phys 2010;107:024508-1–8-8.
[40] Grasser T, Kaczer B. Evidence that two tightly coupled mechanism are responsible for negative bias temperature instability in oxynitride MOSFETs. IEEE Trans Electr Dev 2009;56(5):1056–62.
[41] Hehenberger P, Aichinger T, Grasser T, Goes W, Triebl O, Kaczer B, et al. Do NBTI-induced interface states show fast recovery? A study using a corrected on-the-fly charge-pumping measurement technique. In: Proc IRPS; 2009. p. 1033–8.
[42] Teo Z, Boo A, Ang D, Leong K. On the cyclic threshold voltage shift of dynamic negative-bias temperature instability. In: Proc IRPS; 2011. p. 943–7.
[43] Shockley W, Read W. Statistics of the recombinations of holes and electrons. Phys Rev 1952;87(5):835–42.
[44] Masuduzzaman M, Islam A, Alam M. Exploring the capability of multifrequency charge pumping in resolving location and energy levels of traps within dielectric. IEEE Trans Electr Dev 2008;55(12):3421–31.
[45] Groeseneken G, Maes H, Beltran N, de Keersmaecker R. A reliable approach to charge-pumping measurements in MOS transistors. IEEE Trans Electr Dev 1984;31(1):42–53.
[46] McWhorter A. 1/*f* Noise and germanium surface properties. Sem Surf Phys 1957:207–28.
[47] Henry C, Lang D. Nonradiative capture and recombination by multiphonon emission in GaAs and GaP. Phys Rev B 1977;15(2):989–1016.
[48] Ralls K, Skocpol W, Jackel L, Howard R, Fetter L, Epworth R, et al. Discrete resistance switching in submicrometer silicon inversion layers: individual interface traps and low-frequency (1/*f*?) noise. Phys Rev Lett 1984;52(3):228–31.
[49] Weissman M. 1/*f* Noise and other slow, nonexponential kinetics in condensed matter. Rev Mod Phys 1988;60(2):537–71.
[50] Kirton M, Uren M. Noise in solid-state microstructures: a new perspective on individual defects, interface states and low-frequency (1/*f*) noise. Adv Phys 1989;38(4):367–486.
[51] Campbell J, Qin J, Cheung K, Yu L, Suehle J, Oates A, et al. Random telegraph noise in highly scales nMOSFETs. In: Proc IRPS; 2009. p. 382–8.
[52] Uren M, Kirton M, Collins S. Anomalous telegraph noise in small-area silicon metal–oxide–semiconductor field-effect transistors. Phys Rev B 1988;37(14):8346–50.
[53] Gillespie D. Markov processes: an introduction for physical scientists. Academic Press; 1992.
[54] Ibe O. Markov processes for stochastic modeling. Academic Press; 2009.
[55] Gillespie D. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comput Phys 1976;22:403–34.
[56] Toledano-Luque M, Kaczer B, Simoen E, Roussel P, Veloso A, Grasser T, et al. Temperature and voltage dependences of the capture and emission times of individual traps in high-κ dielectrics. Microelectr Eng 2011;88:1243–6.
[57] Campbell J, Cheung K, Suehle J, Oates A. The fast initial threshold voltage shift: NBTI or high-field stress. In: Proc IRPS; 2008. p. 72–8.
[58] Nicollian E, Brews J. MOS (metal oxide semiconductor) physics and technology. New York: Wiley; 1982.
[59] Asenov A, Balasubramaniam R, Brown A, Davies J. RTS amplitudes in decananometer MOSFETs: 3-D simulation study. IEEE Trans Electr Dev 2003;50(3):839–45.
[60] Kaczer B, Roussel P, Grasser T, Groeseneken G. Statistics of multiple trapped charges in the gate oxide of deeply scaled MOSFET devices-application to NBTI. IEEE Electr Dev Lett 2010;31(5):411–3. doi:10.1109/LED.2010.2044014.
[61] Grasser T, Kaczer B, Goes W, Reisinger H, Aichinger T, Hehenberger P, et al. Recent advances in understanding the bias temperature instability. In: Proc IEDM; 2010. p. 82–5.
[62] Toledano-Luque M, Kaczer B, Roussel P, Franco J, Ragnarsson L, Grasser T, et al. Depth localization of positive charge trapped in silicon oxynitride field effect transistors after positive and negative gate bias temperature stress. Appl Phys Lett 2011;98:183506-1–6-3.
[63] McQuarrie D. Stochastic approach to chemical kinetics. J Appl Prob 1967;4:413–78.
[64] Huard V, Parthasarathy C, Denais M. Single-hole detrapping events in pMOSFETs NBTI degradation. In: IIRW final rep.; 2005. p. 5–9.
[65] Reisinger H, Grasser T, Schlünder C. A study of NBTI by the statistical analysis of the properties of individual defects in pMOSFETs. In: IIRW final rep.; 2009. p. 30–5.
[66] Reisinger H, Grasser T, Gustin W, Schlünder C. The statistical analysis of individual defects constituting NBTI and its implications for modeling DC- and AC-stress. In: Proc IRPS; 2010. p. 7–15.
[67] Toledano-Luque M, Kaczer B, Roussel P, Grasser T, Wirth G, Franco J, et al. Response of a single trap to AC negative bias temperature stress. In: Proc IRPS; 2011. p. 364–71.
[68] Reisinger H, Vollertsen R, Wagner P, Huttner T, Martin A, Aresu S, et al. A study of NBTI and short-term threshold hysteresis of thin nitrided and thick non-nitrided oxides. IEEE Trans Dev Mater Rel 2009;9(2):106–14.
[69] Huang K, Rhys A. Theory of light absorption and non-radiative transitions in F-centres. Proc Roy Soc A 1950;204:406–23.
[70] Fowler W, Rudra J, Zvanut M, Feigl F. Hysteresis and Franck–Condon relaxation in insulator–semiconductor tunneling. Phys Rev B 1990;41(12):8313–7.
[71] Gusmeroli R, Compagnoni C, Riva A, Spinelli A, Lacaita A, Bonanomi M, et al. Defects spectroscopy in SiO_2 by statistical random telegraph noise analysis. In: Proc IEDM; 2006. p. 483–6.
[72] Ghetti A, Compagnoni C, Spinelli A, Visconti A. Comprehensive analysis of random telegraph noise instability and its scaling in deca–nanometer flash memories. IEEE Trans Electr Dev 2009;56(8):1746–52.
[73] Sonoda K, Ishikawa K, Eimori T, Tsuchiya O. Discrete dopant effects on statistical variation of random telegraph signal magnitude. IEEE Trans Electr Dev 2007;54(8):1918–25.
[74] Kaczer B, Grasser T, Roussel P, Franco J, Degraeve R, Ragnarsson L, et al. Origin of NBTI variability in deeply scaled PFETs. In: Proc IRPS; 2010. p. 26–32.
[75] Grasser T, Wagner P-J, Reisinger H, Aichinger T, Pobegen G, Nelhiebel M, et al. Analytic modeling of the bias temperature instability using capture/emission time maps. In: Proc IEDM, in press.
[76] Denais M, Bravaix A, Huard V, Parthasarathy C, Ribes G, Perrier F, et al. On-the-fly characterization of NBTI in ultra-thin gate oxide pMOSFET's. In: Proc IEDM; 2004. p. 109–12.
[77] Shen C, Li M-F, Foo CE, Yang T, Huang D, Yap A, et al. Characterization and physical origin of fast V_{th} transient in NBTI of pMOSFETs with SiON dielectric. In: Proc IEDM; 2006. p. 333–6.
[78] Reisinger H, Brunner U, Heinrigs W, Gustin W, Schlünder C. A comparison of fast methods for measuring NBTI degradation. IEEE Trans Dev Mater Rel 2007;7(4):531–9.
[79] Grasser T, Wagner P-J, Hehenberger P, Goes W, Kaczer B. A rigorous study of measurement techniques for negative bias temperature instability. IEEE Trans Dev Mater Rel 2008;8(3):526–35.
[80] Chakravarthi S, Krishnan A, Reddy V, Machala C, Krishnan S. A comprehensive framework for predictive modeling of negative bias temperature instability. In: Proc IRPS; 2004. p. 273–82.
[81] Mahapatra S, Alam M, Kumar P, Dalei T, Varghese D, Saha D. Negative bias temperature instability in CMOS devices. Microelectr Eng 2005;80(Suppl.):114–21.

[82] Islam A, Mahapatra S, Deora S, Maheta V, Alam M. On the differences between ultra-fast NBTI measurements and reaction–diffusion theory. In: Proc IEDM; 2009. p. 733–6.
[83] Mahapatra S, Maheta V, Islam A, Alam M. Isolation of NBTI stress generated interface trap and hole-trapping components in PNO p-MOSFETs. IEEE Trans Electr Dev 2009;56(2):236–42.
[84] Heiman F, Warfield G. The effects of oxide traps on the MOS capacitance. IEEE Trans Electr Dev 1965;12(4):167–78.
[85] Tewksbury T. Relaxation effects in MOS devices due to tunnel exchange with near-interface oxide traps. Ph.D. thesis. MIT; 1992.
[86] Grasser T, Kaczer B. Negative bias temperature instability: recoverable versus permanent degradation. In: Proc ESSDERC; 2007. p. 127–30.
[87] Grasser T, Kosina H, Selberherr S. Influence of the distribution function shape and the band structure on impact ionization modeling. J Appl Phys 2001;90(12):6165–71.
[88] Gehring A, Grasser T, Kosina H, Selberherr S. Simulation of hot-electron oxide tunneling current based on a non-Maxwellian electron energy distribution function. J Appl Phys 2002;92(10):6019–27.
[89] Lundstrom I, Svensson C. Tunneling to traps in insulators. J Appl Phys 1972;43(12):5045–7.
[90] Schanovsky F, Goes W, Grasser T. Multiphonon hole trapping from first principles. J Vac Sci Technol B 2011;29(1). 01A2011–5.
[91] Campbell J, Lenahan P, Krishnan A, Krishnan S. Observations of NBTI-induced atomic-scale defects. IEEE Trans Dev Mater Rel 2006;6(2):117–22.
[92] Ryan J, Lenahan P, Grasser T, Enichlmair H. Recovery-free electron spin resonance observations of NBTI degradation. In: Proc IRPS; 2010. p. 43–9.
[93] Kimmel A, Sushko P, Shluger A, Bersuker G. Positive and negative oxygen vacancies in amorphous silica. In: Sah R, Zhang J, Kamakura Y, Deen M, Yota J, editors. Silicon nitride, silicon dioxide, and emerging dielectrics 10, vol. 19. ECS Transactions; 2009. p. 2–17.
[94] Kittel C. Introduction to solid-state physics. Wiley; 1996.
[95] Stoneham A. Non-radiative transitions in semiconductors. Rep Prog Phys 1981;44:1251–95.
[96] Fowler W, editor. Physics of color centers. New York: Academic Press; 1968.
[97] Zener C. Non-adiabatic crossing of energy levels. Proc Roy Soc 1932;137:696–702.
[98] Avellán A, Schroeder D, Krautschneider W. Modeling random telegraph signals in the gate current of metal–oxide–semiconductor field effect transistors after oxide breakdown. J Appl Phys 2003;99(1):703–8.
[99] Markvart T. Multiphonon transitions between adiabatic potential curves. J Phys C: Solid State Phys 1984;17:6303–16.
[100] Cheng Z. Unified quantum field theory of light absorption by defect centers. Phys Rev B 1999;23(60):15747–65.
[101] Poindexter E, Warren W. Paramagnetic point defects in amorphous thin films of SiO_2 and Si_3N_4: updates and additions. J Electrochem Soc 1995;142(7):2508–16.
[102] Lelis A, Oldham T. Time dependence of switching oxide traps. IEEE Trans Nucl Sci 1994;41(6):1835–43.
[103] Nicklaw C, Lu Z-Y, Fleetwood D, Schrimpf R, Pantelides S. The structure, properties, and dynamics of oxygen vacancies in amorphous SiO_2. IEEE Trans Nucl Sci 2002;49(6):2667–73.
[104] Blöchl P. First-principles calculations of defects in oxygen-deficient silica exposed to hydrogen. Phys Rev B 2000;62(10):6158–79.