NBTI in Nanoscale MOSFETs—The Ultimate Modeling Benchmark

Tibor Grasser, Karina Rott, Hans Reisinger, Michael Waltl, Franz Schanovsky, and Ben Kaczer

***Abstract*— After nearly half a century of research into the bias temperature instability, two classes of models have emerged as the strongest contenders. One class of models, the reaction-diffusion models, is built around the idea that hydrogen is released from the interface and that it is the diffusion of some form of hydrogen that controls both degradation and recovery. Although various variants of the reaction-diffusion idea have been published over the years, the most commonly used recent models are based on nondispersive reaction rates and nondispersive diffusion. The other class of models is based on the idea that degradation is controlled by first-order reactions with widely distributed (dispersive) reaction rates. We demonstrate that these two classes give fundamentally different predictions for the stochastic degradation and recovery of nanoscale devices, therefore providing the ultimate modeling benchmark. Using detailed experimental time-dependent defect spectroscopy data obtained on such nanoscale devices, we investigate the compatibility of these models with experiment. Our results show that the diffusion of hydrogen (or any other species) is unlikely to be the limiting aspect that determines degradation. On the other hand, the data are fully consistent with reaction-limited models. We finally argue that only the correct understanding of the physical mechanisms leading to the significant device-to-device variation observed in the degradation in nanoscale devices will enable accurate reliability projections and device optimization.**

***Index Terms*— Bias temperature instability, charge trapping, dispersive reaction rates, first-order processes, NBTI, oxide defects, PBTI, reaction-diffusion.**

I. Introduction

RESEARCH into the bias temperature instability (BTI) has revealed a plethora of puzzling issues which have proven a formidable obstacle to the understanding of the phenomenon [1]–[14]. In particular, numerous modeling ideas have been put forward and refined at various levels. Most of these models have in common that the overall degradation is assumed to be caused by two components: one component ($N_{\rm it}$) is related to the release of hydrogen from passivated silicon dangling bonds at the interface, thereby forming electrically active $P_{\rm b}$ centers [15], and the other ($N_{\rm ot}$) is owing to the trapping of holes in the oxide [5], [8], [16]–[19]. However, these models can differ significantly in the details of the physical mechanisms invoked to explain the degradation.

At present, from all these modeling attempts two classes have emerged that appear to be able to explain a wide range of experimental observations: the first class is built around the concept of the reaction-diffusion (RD) model [1], [14], where it is assumed that it is the diffusion of the released hydrogen that dominates the dynamics. The other class is based on the notion that it is the reactions which essentially limit the dynamics, and that the reaction rates are distributed over a wide range [5], [20]–[23]. In other words, in this reaction-limited class of models, both interface states ($N_{\rm it}$) and oxide charges ($N_{\rm ot}$) are assumed to be (in the simplest case) created and annealed by the first-order reactions. In contrast, in the diffusion-limited class (RD models), the dynamics of $N_{\rm it}$ creation and annealing are assumed to be dominated by a diffusion-limited process, which controlles both long-term degradation and recovery.

Many of these models have been developed to such a high degree that they appear to be able to predict a wide range of experimental observations [9], [12], [14], [24], [25]. Typically, however, experimental data are obtained on large-area (macroscopic) devices where the microscopic physics are washed out by averaging. In nanoscale devices, on the other hand, it has been shown that the creation and annihilation of individual defects can be observed at the statistical level [3], [5], [10], [13], [24]. We will demonstrate in the following that this statistical information provides the ultimate benchmark for any BTI model, as it reveals the underlying microscopic physics to an unprecedented degree. This allows for an evaluation of the foundations of the two model classes, as it clearly answers the fundamental question: is BTI reaction- or diffusion-limited? As such, the benchmark provided here is simple and not clouded by the complexities of the individual models.

II. Equivalence of Large and Small Devices

Because the stochastic response of nanoscale devices to bias-temperature stress lies at the heart of our arguments, we begin by experimentally demonstrating the equivalence of large- and small-area devices. Therefore, we compare the degradation of a large-area device to the average degradation observed in 27 small-area devices when subjected to negative BTI (NBTI). All measurements in this paper rely on the

Manuscript received June 5, 2014; revised August 5, 2014; accepted August 22, 2014. Date of publication September 18, 2014; date of current version October 20, 2014. This work was supported by the Austrian Science Fund under Project 23390-M24 and in part by the European Community's FP7 Programme under Grant 261868 through the MORDRED Project. The review of this paper was arranged by Editor E. Rosenbaum.

T. Grasser, M. Waltl, and F. Schanovsky are with the Institute for Microelectronics, Technische Universität Wien, Vienna 1040, Austria (e-mail: grasser@iue.tuwien.ac.at; waltl@iue.tuwien.ac.at; schanovsky@iue.tuwien.ac.at).

K. Rott and H. Reisinger are with Infineon, Munich 85579, Germany (e-mail: karina.rott@infineon.com; hans.reisinger@infineon.com).

B. Kaczer is with imec, Leuven 3001, Belgium (e-mail: kaczer@imec.be).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2014.2353578

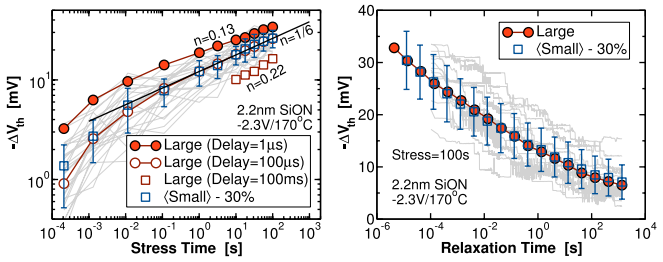

Fig. 1. Degradation (left) and recovery (right) of 27 small-area devices (light-gray lines) (120 nm×280 nm) compared with a large-area device (red symbols) with area 120 nm × 10 µm. Although the average degradation of the small-area devices is larger by 30% (open symbols, error bars are $\pm\sigma$), the kinetics during both stress and recovery are otherwise identical. In particular, during stress a power-law slope of 1/6 is observed in both large and small-area devices, if the measurement delay is chosen 100 µs.

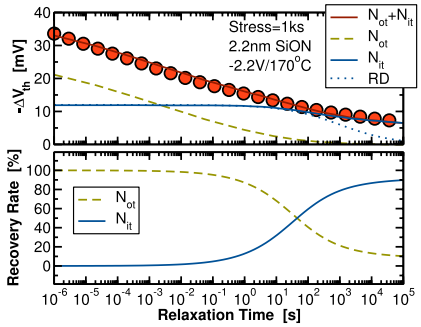

Fig. 2. Top: Recovery data (symbols) from our technology [23] after 1 ks fitted by a simple hole-trapping model ($N_{\rm ot}$) and the empirically modified RD model ($N_{\rm it}$), as taken from [27]. Dotted line (RD): Prediction of the unmodified RD model. Bottom: After about 50 s, according to this fit, recovery is dominated by reaction-limited $N_{\rm it}$ recovery. The recovery rate R is defined by how much $\Delta V_{\rm th}$ is lost per decade in percent.

ultrafast $\Delta V_{\rm th}$ technique published previously [4], which has a delay of 1 µs on large devices. Because of the lower current levels, the delay increases to 100 µs in small-area devices. As shown in Fig. 1, although the degradation in small-area devices shows larger signs of variability, discrete steps during recovery, and is about 30% larger than in this particular large-area device, the average dynamics are identical [10], [26]. In particular, for a measurement delay of 100 µs, a power-law in time ($t_{\rm s}^n$) with exponent 1/6 is observed during stress while the averaged recovery is roughly logarithmic over the relaxation time $t_{\rm r}$. This demonstrates that by using nanoscale devices, the complex phenomenon of NBTI can be broken down to its microscopic constituents: the defects that cause the discrete steps in the recovery traces. Analysis of the statistics of these steps will thus reveal the underlying physical principles.

It has been shown that the hole-trapping component depends sensitively on the process details, particularly for high nitrogen contents [14], possibly making the choice of benchmark technology crucial for our following arguments. However, for industrial grade devices with low nitrogen content such as those used in this paper, no significant differences in reported $\Delta V_{\rm th}$ drifts to published data have been found [10]. The pMOS samples used here are from a standard 120 nm CMOS process with a moderate oxide thickness of 22 Å and with a nitride content of approximately 6%, whereas the poly-Si gates are boron doped with a thickness of 150 nm. In particular, our previously published data obtained on the same technology as that of Fig. 1 have recently been interpreted from the RD perspective [27] as shown in Fig. 2, without showing any anomalies. This fit seems to suggest that after $t_{\rm s} = 1$ ks and $t_{\rm r} > 50$ s recovery is dominated by diffusion-limited $N_{\rm it}$ recovery, a conclusion we will put to the test in the following.

III. Experimental Method

For our experimental assessment, we use time-dependent defect spectroscopy (TDDS) [24], which has been extensively used to study BTI in small-area devices at the single-defect level [3], [13], [28], [29]. Because such devices contain only a countable number of defects, the recovery of each defect is visible as a discrete step in the recovery trace (Fig. 1). The large variability of the discrete step-heights is a consequence of the inhomogeneous surface potential caused by the random discrete dopants in the channel, leading to percolation paths and a strong sensitivity of the step-height to the spatial location of the trapped charge [30]. Typically, these step-heights are approximately exponentially distributed [31] with the mean step-height given by $\bar{\eta} = \bar{\eta}_{\rm r}\eta_0$. Here, η_0 is the value expected from the simple charge sheet approximation $\eta_0 = qt_{\rm ox}/(\epsilon_{\rm r}\epsilon_0 WL)$, where q is the elementary charge, $\epsilon_{\rm r}\epsilon_0$ is the permittivity of the oxide, WL is the area, and $t_{\rm ox}$ is the oxide thickness. Experiments and theoretical values for the mean correction factor $\eta_{\rm r}$ are in the range 1–4 [32].

In a TDDS setup, a nanoscale device is repeatedly stressed and recovered (e.g., $N = 100$ times) using fixed stress/recovery times, $t_{\rm s}$ and $t_{\rm r}$. The recovery traces are analyzed for discrete steps of height η occurring at time $\tau_{\rm e}$. Each $(\tau_{\rm e}, \eta)$ pair is then placed into a 2-D histogram, which we call the spectral map, formally denoted as $g(\tau_{\rm e}, \eta)$. The clusters forming in the spectral maps reveal the probability density distribution, and thus provide detailed information on the statistical nature of the average trap annealing time constant $\bar{\tau}_{\rm e}$. From the evolution of $g(\tau_{\rm e}, \eta)$ with stress time, the average capture time $\bar{\tau}_{\rm c}$ can be extracted as well. So far, only exponential distributions have been observed for $\tau_{\rm e}$, consistent with simple independent first-order reactions [33].

In our previous TDDS studies, mostly short-term stresses ($t_{\rm s} \lesssim 1$ s) had been used. With this short-term nature, the generality of these results may be questioned, because also $N_{\rm it}$ recovery predicted by RD models result in discrete steps [34]. As we have pointed out a while ago [35], the distribution of these RD steps would, however, be loglogistic rather than exponential, a fact that should be clearly visible in the spectral maps. In the following, we will conduct a targeted search for such loglogistic distributions and other features directly linked to diffusion-limited recovery processes by using extended long-term TDDS experiments with $t_{\rm s} = t_{\rm r} = 1$ ks.

IV. Theoretical Predictions

Before discussing the long-term TDDS data, we summarize the basic theoretical predictions of the two model classes.

Both model classes have in common that the charges trapped in interface and oxide states induce a change of the threshold voltage. Depending on the location of the charge along the interface or in the oxide, it will contribute a discrete step η_i to the total $\Delta V_{\rm th}$. Because of only occasional electrostatic interactions with other defects and measurement noise, η_i is typically normally distributed with mean $\bar{\eta}_i$. The mean values $\bar{\eta}_i$ themselves, however, are exponentially distributed [31].

The major difference between the model classes is whether creation and annealing of $N_{\rm it}$ is diffusion- or reaction-limited, resulting in a fundamentally different form of the spectral map $g(\tau_{\rm e}, \eta)$, as will be derived below. Considering the simpler case, we begin with the dispersive reaction-limited models.

A. Dispersive Reaction-Limited Models

In an agnostic formulation of dispersive reaction-limited models, creation and annealing of a single defect are assumed to be given by a simple first-order reaction

$$f(t_{\rm s}, t_{\rm r}, \bar{\tau}_{\rm c}, \bar{\tau}_{\rm e}) = (1 - \exp(-t_{\rm s}/\bar{\tau}_{\rm c}))\exp(-t_{\rm r}/\bar{\tau}_{\rm e}) \quad (1)$$

where f is the probability of having a charged defect after stress and recovery times $t_{\rm s}$ and $t_{\rm r}$, respectively. The physics of trap creation enter the average forward and backward time constants $\bar{\tau}_{\rm c}$ and $\bar{\tau}_{\rm e}$. It is important to highlight that (1) may describe both the reaction-limited creation and annealing of interface states [5], [20], [36], as well as a charge-trapping process [3], [5], [24]. We recall that even more complicated charge-trapping processes involving structural relaxation and metastable defect states (such as switching oxide traps) can be approximately described by an effective first-order process, at least under quasi-dc conditions [33], [37].

Having N defects present in a given device, the overall $\Delta V_{\rm th}$ is then simply given by a sum of such first-order processes

$$\Delta V_{\rm th}(t_{\rm s}, t_{\rm r}) = \sum_i \bar{\eta}_i f(t_{\rm s}, t_{\rm r}, \bar{\tau}_{{\rm c},i}, \bar{\tau}_{{\rm e},i}). \quad (2)$$

The most important aspect is that the time constants are observed to be widely distributed. We have recently used such a model to explain BTI degradation and recovery over a very wide experimental window assuming the time constants to belong to two different distributions, one tentatively assigned to charge-trapping and the other to interface state generation [23], [26].

At the statistical level, recovery in such a model is described by the sum of exponential distributions. The spectral map, which records the emission times on a logarithmic scale, is then given by

$$g(\tau_{\rm e}, \eta) = \sum_i B_i \, f_\eta\left(\frac{\eta - \bar{\eta}_i}{\sigma_{\eta,i}}\right)\frac{t_{\rm r}}{\bar{\tau}_{{\rm e},i}}\exp(-t_{\rm r}/\bar{\tau}_{{\rm e},i}) \quad (3)$$

with the stress time-dependent amplitude $B_i \approx 1 - \exp(-t_{\rm s}/\bar{\tau}_{{\rm c},i})$ and f_η describing the p.d.f. of η, with mean $\bar{\eta}_i$ and standard deviation $\sigma_{\eta,i}$. An example spectral map simulated at two different stress times is shown in Fig. 3, which clearly reveals the three contributing defects. We note already here that contrary to the RD model, the spectral map of the dispersive first-order model depends on the individual $\bar{\tau}_{{\rm e},i}$, which can be strongly bias and temperature dependent.

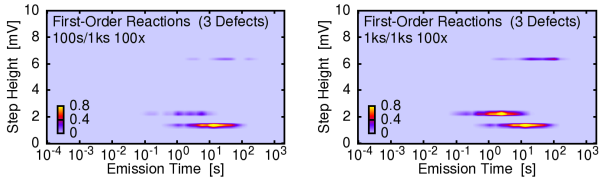

Fig. 3. Simulated spectral maps of a dispersive reaction model for three traps using two stress times: 100 s and 1 ks (left versus right). The map is constructed using 100 repeated stress or relax cycles. The basic features are exponential clusters which do not move with stress time.

B. Nondispersive Reaction-Diffusion Models

As a benchmark RD model we take the latest, and according to [14] the physically most likely variant, the poly H/H_2 model: here it is assumed that H is released from Si–H bonds at the interface, diffuses to the oxide–poly interface, where additional Si–H bonds are broken to eventually create H_2, the diffusion of which results in the $n = 1/6$ degradation behavior typically associated with RD models. Recovery then occurs via reversed pathways. Although other variants of the RD model have been used [1], [38]–[40], which cannot possibly be exhaustively studied here, we believe our findings are of general validity, as all these models are built around diffusion-limited processes.

In large-area devices, the predicted long-term recovery after long-term stress can be fitted by the empirical relation [7]

$$N_{\rm it}(t_{\rm s}, t_{\rm r}) \approx \frac{A t_{\rm s}^n}{1 + (t_{\rm r}/t_{\rm s})^{1/s}} \quad (4)$$

with $s \approx 2$, provided diffusion is allowed into a semi-infinite gate-stack with constant diffusivity to avoid saturation effects. Quite intriguingly, a similar mathematical form has been successfully used to fit a wide range of experimental data, using a scaled stress time [7]. Remarkably, experimentally observed exponents $1/s$ are considerably smaller than what is predicted by RD models, corresponding to a wider spread over the time axis.

In an empirically modified model, it has been assumed that in a real 3-D device, recovery will take longer compared with (4), because the H atoms will have to hover until they can find a suitable dangling bond for passivation [14]. However, using a rigorous stochastic implementation of the RD model, we have not been able to observe significant deviations from (4), irrespective of whether the model is solved in 1-D, 2-D, or 3-D, provided one is in the diffusion-limited regime [41]. As such, significant deviations from the basic recovery behavior (4) still have to be rigorously justified. One option to stretch the duration of recovery would be the consideration of dispersive transport [42], [43]. Our attempts in this direction were, however, not found to be in agreement with experimental observations [7], [12]. Alternatively, consistent with experiment [20], a distribution in the forward and backward reactions can be introduced into the model [40]. This dispersion will stretch the distribution (4), that is, increase the parameter s, but may also lead to a temperature dependence of

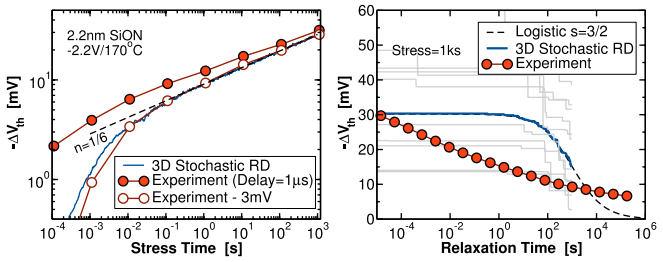

Fig. 4. Degradation (left) and recovery (right) predicted by the calibrated stochastic poly H/H_2 model on small-area devices. The difference in the initial stress phase is assumed to be because of hole trapping and approximately modeled by subtracting 3 mV from the experimental data, as we are here concerned with larger stress and recovery times, where hole trapping is assumed to be negligible in the RD interpretation [44]. Recall that Fig. 2 uses the empirically stretched RD model [27].

the power-law slope, features which have not been validated so far. Nevertheless, a dispersion in the reaction-rates as used for instance in [40] will not change the basic diffusion-limited nature of the microscopic prediction as shown in the following.

To study the stochastic response of the poly H/H_2 model, we extended our previous stochastic implementation [41] of the H/H_2 RD model to include the oxide/poly interface following ideas and parameters of [45]. Because any sensible macroscopic model is built around a well-defined microscopic picture, in this case nondispersive diffusion and nondispersive rates, these features of the microscopic model must be preserved in the macroscopic theory, leaving little room for interpretation. To be consistent with the $W \times L = 150\,\text{nm} \times 100\,\text{nm}$ devices used in our TDDS study, we chose $\bar{\eta} = \bar{\eta}_r \eta_0 = 2 \times 0.9\,\text{mV} = 1.8\,\text{mV}$ [24]. Furthermore, a typical density of interface states $N_{it} = 2 \times 10^{12}\,\text{cm}^{-2}$ [20], [40] is assumed. We would thus expect about 300 such interface states to be present for our TDDS devices.

Before looking into the predictions of this RD model in a TDDS setting, we calibrate our implementation of the poly H/H_2 model to experimental stress data [Fig. 4 (left)]. To obtain a good fit during stress, we follow the procedure suggested in [44] and subtract a virtual hole-trapping contribution of 3 mV from the experimental data to obtain the required $n = 1/6$ power-law. Also, we remark that to achieve this fit, unphysically large hydrogen hopping distances had to be used in the microscopic model [41]. Furthermore, H_2 had to be allowed to diffuse more than a micrometer deep into the gate-stack with unmodified diffusion constant to maintain the $n = 1/6$ power-law exponent, despite the fact that our poly-Si gate was only 150 nm thick.

From (4), we can directly calculate the expected unnormalized probability density function for RD recovery as [33]

$$f_{\text{RD}}(t_r) = -\frac{\partial N_{it}(t_r)}{\partial \log(t_r)} = A t_s^n \frac{(t_r/t_s)^{1/s}}{s\left(1+(t_r/t_s)^{1/s}\right)^2} \tag{5}$$

which after normalization by $A t_s^n$ is a loglogistic distribution of $\log(t_r)$ with parameter s and mean $\log(t_s)$. In the framework of the standard nondispersive RD model, all interface states are equivalent in the sense that on average they will have degraded and recovered with the same probability at a certain stress or recovery time combination. In terms of impact on ΔV_{th},

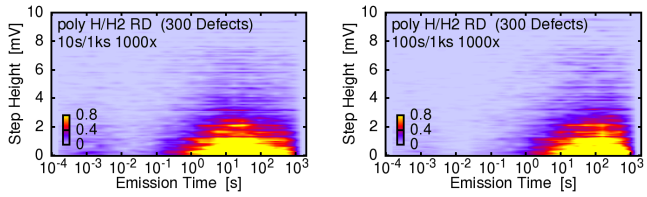

Fig. 5. Since in nondispersive RD models all defects contribute equally to the spectral map, no clear clusters can be identified, except for possibly in the tail of the exponential distribution. Shown is a poly H/H_2 simulation with 300 defects for two stress times. Note that on average all defects are active with the same probability at all times, which results in markedly different spectral maps compared with those produced by a dispersive reaction model (Fig. 3).

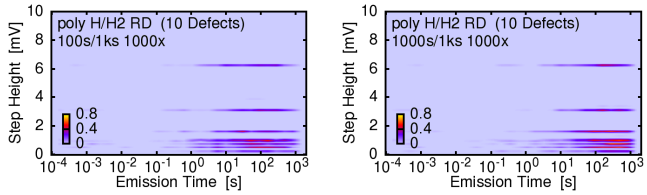

Fig. 6. Simulated spectral maps using the poly H/H_2 model for a 20 nm × 25 nm sized device with only ten active defects. Note how the intensity, that is, the emission probability, of the clusters keeps increasing while the mean of the clusters shifts to larger times with increasing stress time.

we again assume that the mean impact of a single trap $\bar{\eta}_i$ is exponentially distributed.

Using (5), the spectral map built of subsequent stress or relax cycles can be obtained. Because all defects are controlled by the same diffusive process and are thus equivalent except for their step-heights, the time dynamics can be pulled out of the sum to eventually give

$$g(\tau_e, \eta) = A t_s^n \frac{(t_r/t_s)^{1/s}}{s\left(1+(t_r/t_s)^{1/s}\right)^2} \sum_i f_\eta\left(\frac{\eta - \bar{\eta}_i}{\sigma_{\eta,i}}\right). \tag{6}$$

This is a very interesting result, as it implies that all defects are active with the same probability at any time, leading to a dense response in the spectral map as shown in Fig. 5. As will be shown, this is incompatible with our experimental results.

To more clearly elucidate the features of the RD model, in the following we will use a 20 nm × 25 nm device, in which only a small number of defects (about ten) contribute to the spectral maps. The crucial fingerprint of the RD model would then be that these clusters are loglogistically distributed and thus much wider than the previously observed exponential distributions. Furthermore, we note that the RD spectral map does not depend on any parameter of the model nor does it depend on temperature and bias [1], but owing to the diffusion-limited nature of the model shifts to larger times with increasing stress time (Fig. 6), facts we will compare against experimental data later.

V. Small-Devices: Purely Reaction-Limited

As noted before, previous TDDS experiments had been limited to stress times mostly smaller than about 1 s, which may limit the relevance of our findings for long-term stress. As such, it was essential to extend the stress and relaxation

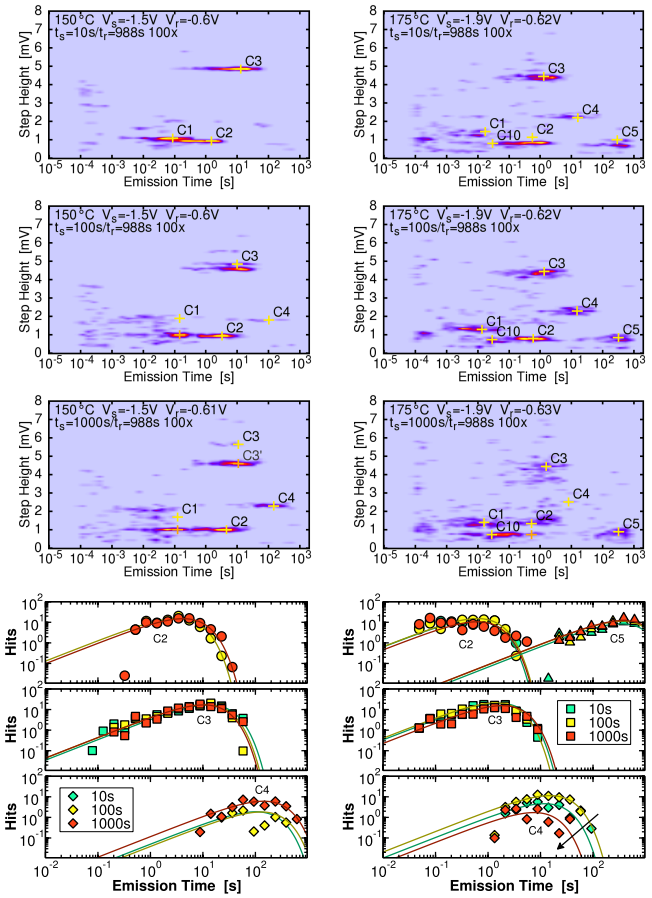

Fig. 7. Even at longer stress times (10 s – 1 ks) and higher temperatures, 150 °C (left/top) and 175 °C (right/top), all clusters (symbols) are exponential (lines) and do not move with stress time (bottom), just like the prediction of a reaction-limited model (Fig. 3). Because of the increasing number of defects contributing to the emission events, the data become noisier with increasing stress bias, temperature, and time. With increasing stress, defect C4 shows signs of volatility, leading to a smaller number of emission events at longer times [49].

times to 1 ks, which is a typically used experimental window [14]. Unfortunately, the stress or relax cycles needed to be repeated at least 100 times, otherwise differentiation between exponential and logistic distributions would be difficult. We therefore used nine different stress times for each experiment, starting from 10 µs up to 1 ks with recovery lasting 1 ks, repeated 100 times, requiring a total of about 12 days. About 20 such experiments were carried out on four different devices over the course of more than half a year.

Because we are particularly interested in identifying a diffusion-limited contribution to NBTI recovery, we tried to minimize the contribution of charge trapping. With increasing stress voltage, an increasing fraction of the bandgap becomes accessible for charging [33], therefore we primarily used stress voltages close to $V_{DD} = -1.5$ V of our technology (about 4 MV/cm [46]). Furthermore, it has been observed that at higher stress voltages, defect generation in a TDDB-like degradation mode can become important [14], [47], [48], an issue we avoid at such low stress voltages. Two example measurements are shown in Fig. 7 for −1.5 V at 150 °C and −1.9 V at 175 °C (about 4–5 MV/cm). As already observed for short-term stresses, all clusters

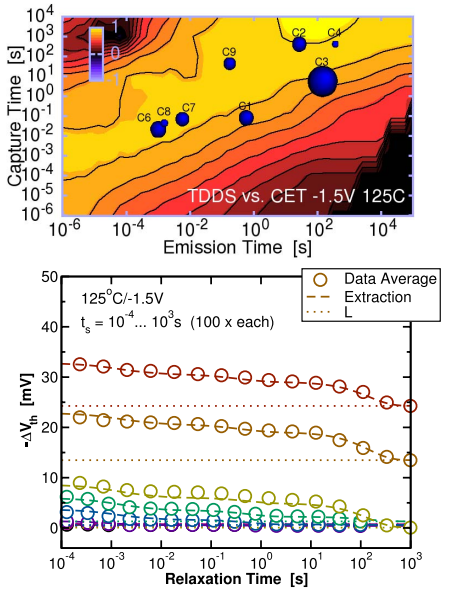

Fig. 8. Top: Comparison of the extracted capture and emission times versus a CET map from a large area device [23]. Size of the dots: η value of each defect. The distribution of the individual defects seen in TDDS agrees well with the CET map. Bottom: Using the time constants extracted from the long-term TDDS data (lines), it is possible to fully reconstruct the average recovery traces (symbols, corresponding to the expectation value) for all stress and recovery times. The average offset L at $t_r = 1$ ks is added (dotted lines) to show the buildup of defects with larger emission or annealing times (∼permanent component).

are exponential and have a temperature-dependent but time-independent mean $\bar{\tau}_e$. Most noteworthy is the fact that no sign of an RD signature as discussed in Section IV-B was observed. We remark that defects tend to show strong signs of volatility at longer stress and recovery times [49], a fascinating issue to be discussed in more detail elsewhere.

To confirm that our extracted capture and emission times fully describe recovery on average, we calculate the average of all 100 recovery traces recorded at each stress time and compare it with the prediction given by the extracted $\bar{\tau}_{c,i}$ and $\bar{\tau}_{e,i}$ values using (2), which corresponds to the expectation value and thus the average. Indeed, as shown in Fig. 8, excellent agreement is obtained, finally proving that our extraction captures the essence of NBTI recovery. It is worth pointing out that this agreement is obtained without fitting of the average data: we simply use the extracted capture and emission times as well as the extracted step-heights and put them into (2). Also shown is a comparison of the capture or emission times extracted by TDDS with a capture or emission time (CET) map extracted on large devices [23]. The capture and emission times extracted on the nanoscale device are fully consistent with the macroscopic distribution and correspond to a certain realization, which will vary from device to device.

As a final point, we compare the averaged recovery over 100 repetitions obtained from four different nanoscale devices after 1 ks stress under the same conditions as shown in Fig. 9. Clearly, all devices recover in a very unique way. For instance, device F shows practically no recovery between 10 s and 1 ks, while device D has a very strong recoverable component in this time window but practically no recovery from 1 ms up to 10 s. Furthermore, this unique recovery

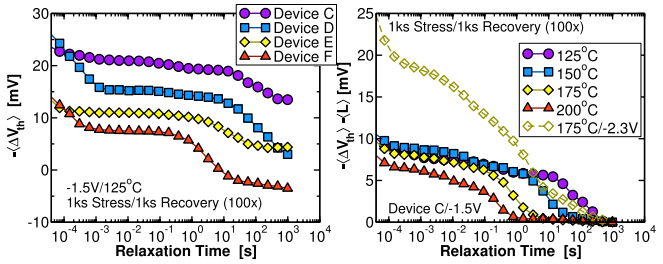

Fig. 9. Left: Just like for short-term stress, the averaged recovery over 100 repetitions after long-term stress or relax cycles varies strongly from device to device. Also noteworthy is the dramatic difference in the last value at $t_r = 1$ ks of each averaged recovery trace, meaning that also the buildup of the permanent component is stochastic. Right: Contrary to the RD prediction, recovery depends strongly on temperature and stress bias. Shown is the average recovery data minus the averaged last value of each recovery trace $\langle L \rangle$. In device C, because of the strong temperature activation of the emission time constants, the average ΔV_{th} traces are shifted to shorter times with increasing T, leading to practically no recovery after 1 s for $V_s = -1.5$ V. At higher stress voltages ($V_s = -2.3$ V), a considerably larger number of traps (presumably in the oxide) can contribute, leading to strong recovery in the whole measurement window.

depends strongly on bias and temperature, as shown in Fig. 9 (right) for device C. For example, after a stress at -1.5 V at 125 °C, strong recovery is observed between 10 s and 1 ks, which is completely absent at 200 °C. On the other hand, if the stress bias is increased to say -2.3V (about 7 MV/cm), a nearly logarithmic recovery is observed in the whole experimental window, consistent with what is also seen in large-area devices.

In the nondispersive RD picture, hundreds of defects would be equally contributing to the average recovery of such devices. As such, the model is practically immune to the spatial distribution of the defects which would be the dominant source of device-to-device variability in this nondispersive RD picture, lacking any other significant parameters. Such a model can therefore not explain the strong device-to-device variations observed experimentally. Also, as discussed before, in nondispersive RD models in their present form recovery is independent of bias and temperature, which is also at odds with these data.

On the other hand, our data is perfectly consistent with a collection of defects with randomly distributed $\bar{\tau}_{c,i}$ and $\bar{\tau}_{e,i}$. In this picture, the occurrence of a recovery event only depends on whether a defect with a suitable pair ($\bar{\tau}_{e,i}$, $\bar{\tau}_{c,i}$) exists in this particular device. As these time constants depend on bias and temperature, the behavior seen in Fig. 9 is a natural consequence. Three additional finer points related to our extraction and analysis are discussed in the Appendix.

VI. Consequences

The question whether NBTI is due to a diffusion- or reaction-limited process is of high practical significance and not merely a mathematical modeling detail. First of all, it is essential from a process optimization point of view: if the RD model in any variant were correct, then one should seek to prevent the diffusion into the gate-stack by, for instance, introducing hydrogen diffusion barriers. This is because according to RD models, upon hitting such a barrier, the hydrogen concentration in the gate-stack would equilibrate, leading to an end of the degradation. On the other hand, if reaction-limited models are correct—and our results clearly indicate that they are—device optimization from a reliability perspective should focus on the distribution of the time constants or reaction rates in the close vicinity of the channel that are responsible for charge trapping and the reaction-limited creation of interface states.

Second, our results have a fundamental impact on our understanding of the stochastic reliability of nanoscale devices. We have demonstrated that even the averaged response of individual devices will be radically different from device to device, whereas in nondispersive RD models all devices will on average degrade in the same manner. Given the strong bias- and temperature-dependence of this individual response, it is mandatory to study and understand the distribution of the bias- and temperature-dependence of the responsible reaction rates. This is exactly the route taken recently in [50], where it was shown that the energetic alignment of the defects in the oxide with the channel can be tuned by modifying the channel materials to optimize device reliability.

VII. Conclusions

Using nanoscale devices, we have established an ultimate benchmark for BTI models at the statistical level. Contrary to previous studies, we have used a very wide experimental window, covering stress, and recovery times from the microsecond regime up to kiloseconds, as well as temperatures up to 175 °C. The crucial observations are as follows.

1) Using TDDS, all recovery events create exponentially distributed clusters on the spectral maps which do not move with increasing stress time.
2) The location of these clusters is marked by a capture time, an emission time, and the step-height. In an agnostic manner, we also consider the forward and backward rates for the creation of interface states on the same footing. The combination of such clusters forms a unique fingerprint for each nanoscale device.
3) Given the strong bias- and temperature-dependence of the capture and emission times, the degradation in each device will have a unique temperature- and bias-dependence.

At the microscopic level, any BTI model describing charge trapping as well as the creation of interface states should be consistent with the preceding findings. From the wide variety of published models, we have compared two model classes with these benchmarks, namely reaction- versus diffusion-limited models.

As a representative for diffusion-limited models, we have used the poly H/H_2 RD model. We have observed a complete lack of agreement, as this nondispersive RD model predicts that: 1) a very large number of equal interface states contribute equally to recovery, whereas experimentally only a countable number of clusters can be identified; 2) the clusters observed in the spectral map should be loglogistically distributed with an increasing mean value given by the stress-time; and 3) the averaged long-term degradation and recovery should be roughly the same in all devices, independent of temperature

and bias. With these observations, we conclude that the mainstream nondispersive RD models in their present form are unlikely to provide a correct physical picture of NBTI. These issues should be addressed in future variants of RD models and benchmarked against the observations made here.

On the other hand, if we go to the other extreme and assume that NBTI recovery is not diffusion-limited but reaction-limited, the characteristic experimental signatures are naturally reproduced. Such models: 1) are consistent with the exponential distributions in the spectral map; 2) are based on widely dispersed capture and emission times which result in fixed clusters on the spectral maps; and 3) naturally result in a unique fingerprint for each device, as the parameters of the reaction are drawn from a wide distribution. As the time constants are bias- and temperature-dependent, the unique behavior of each device can be naturally explained and predicted, provided the distribution of these time constants is understood.

Finally, we have argued that our results are not only interesting for modeling enthusiasts, but have fundamental practical implications regarding the way devices should be optimized and analyzed for reliability, particularly for nanoscale devices, which will show increased variability.

APPENDIX

In this appendix, three finer points are discussed, namely: 1) the subtle difference between fully independent stress/relax cycles implied by (5) and a TDDS setting; 2) a possible impact of errors in the discrete-step extraction algorithm; and 3) a contribution of the quasi-permanent component.

A. Repeated Stress or Relax Cycles

Strictly speaking, (5) is valid for a single stress or relax cycle while the TDDS consists of a large number of repeated cycles. As such, the TDDS setup corresponds to an ultralow-frequency ac stress and the devices will not be fully recovered prior to the next stress phase. This implies that H would be able to move deeper into the gate-stack during cycling and that the H profile would not be precisely the same as that predicted during dc stress [11]. For short stress times and long enough recovery times, for example, 1 s versus 1 ks, the impact of this would be small, because (5) predicts nearly full recovery in this case (97%). However, for larger stress times, recovery by the end of the cycle will only be partial and (5) may no longer be accurate in a TDDS setting. We have considered this case numerically in Fig. 10 (left), which shows that although this impacts the absolute number of recorded emission events, the general features—namely loglogistically distributed clusters which move in time– remain.

B. On Possible Extraction Errors

As can be seen from Fig. 7, with increasing stress time the number of visible clusters increases, as does the noise-level, making an accurate extraction of the statistical parameters more challenging than for shorter stress times. To guarantee that our extraction algorithm, which splits the recovery trace into discrete steps, does not miss any essential features and the noise in the spectral maps is really just unimportant noise

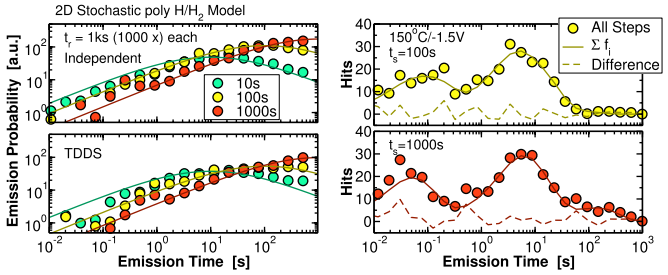

Fig. 10. Left: (Top) Emission probability predicted by the poly H/H_2 model (symbols are simulation) assuming independent stress or relax cycles follows a loglogistic distribution (lines). (Bottom) If the simulations are not conducted independently (i.e., repeated on a completely recovered device) but in a TDDS setting (1000 repeated stress or relax cycles for better statistics), the number of emission events decreases but the statistics remain nearly unaffected. Right: The sum of the exponential distributions fitted to the individual clusters (lines) is subtracted from the total number of detected switches (symbols), revealing a certain noise in the data. However, no hidden RD component is identifiable in the noise.

rather than an overshadowed RD contribution, we performed one additional test: we calculate the difference between the extracted response of forward and backward reactions and subtract it from all recorded steps [Fig. 10 (right)]. As can be seen, even if owing to noise not all steps are considered in the fit, no hidden RD component is missed.

C. Permanent Contribution to TDDS

Finally, we comment on the permanent part that builds up during the TDDS cycles (Fig. 8). This contribution is not explicitly modeled here, but only extracted from the experimental data to be added to the modeled recoverable part. From an agnostic perspective, one could simply refer to this permanent buildup as that due to those defects with emission or annealing times larger than the maximum recovery time, 1 ks in our case. This permanent buildup is typically assigned to interface states (P_b centers) [5], but likely also contains a contribution from charge traps with large time constants [51].

REFERENCES

[1] M. A. Alam, "A critical examination of the mechanics of dynamic NBTI for pMOSFETs," in *Proc. Int. Electron Devices Meeting (IEDM)*, Dec. 2003, pp. 345–348.

[2] B. Kaczer, V. Arkhipov, R. Degraeve, N. Collaert, G. Groeseneken, and M. Goodwin, "Temperature dependence of the negative bias temperature instability in the framework of dispersive transport," *Appl. Phys. Lett.*, vol. 86, no. 14, pp. 143506-1–143506-3, Apr. 2005.

[3] T. Wang *et al.*, "A novel transient characterization technique to investigate trap properties in HfSiON gate dielectric MOSFETs-from single electron emission to PBTI recovery transient," *IEEE Trans. Electron Devices*, vol. 53, no. 5, pp. 1073–1079, May 2006.

[4] H. Reisinger, O. Blank, W. Heinrigs, A. Mühlhoff, W. Gustin, and C. Schlünder, "Analysis of NBTI degradation- and recovery-behavior based on ultra fast V_T-measurements," in *Proc. 44th Annu. IEEE Int. Rel. Phys. Symp. (IRPS)*, Mar. 2006, pp. 448–453.

[5] V. Huard, M. Denais, and C. Parthasarathy, "NBTI degradation: From physical mechanisms to modelling," *Microelectron. Rel.*, vol. 46, no. 1, pp. 1–23, 2006.

[6] M. Houssa, V. V. Afanas'ev, A. Stesmans, M. Aoulaiche, G. Groeseneken, and M. M. Heyns, "Insights on the physical mechanism behind negative bias temperature instabilities," *Appl. Phys. Lett.*, vol. 90, no. 4, pp. 043505-1–043505-3, Jan. 2007.

[7] T. Grasser, W. Gos, V. Sverdlov, and B. Kaczer, "The universality of NBTI relaxation and its implications for modeling and characterization," in *Proc. 45th Annu. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2007, pp. 268–280.

[8] J. F. Zhang, Z. Ji, M. H. Chang, B. Kaczer, and G. Groeseneken, "Real V_{th} instability of pMOSFETs under practical operation conditions," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2007, pp. 817–820.

[9] V. Huard, "Two independent components modeling for negative bias temperature instability," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, May 2010, pp. 33–42.

[10] H. Reisinger, T. Grasser, W. Gustin, and C. Schlünder, "The statistical analysis of individual defects constituting NBTI and its implications for modeling DC- and AC-stress," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, May 2010, pp. 7–15.

[11] D. S. Ang, Z. Q. Teo, T. J. J. Ho, and C. M. Ng, "Reassessing the mechanisms of negative-bias temperature instability by repetitive stress/relaxation experiments," *IEEE Trans. Device Mater. Rel.*, vol. 11, no. 1, pp. 19–34, Mar. 2011.

[12] T. Grasser *et al.*, "The paradigm shift in understanding the bias temperature instability: From reaction–diffusion to switching oxide traps," *IEEE Trans. Electron Devices*, vol. 58, no. 11, pp. 3652–3666, Nov. 2011.

[13] J. Zou *et al.*, "New insights into AC RTN in scaled high-κ/metal-gate MOSFETs under digital circuit operations," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2012, pp. 139–140.

[14] S. Mahapatra *et al.*, "A comparative study of different physics-based NBTI models," *IEEE Trans. Electron Devices*, vol. 60, no. 3, pp. 901–916, Mar. 2013.

[15] J. P. Campbell, P. M. Lenahan, C. J. Cochrane, A. T. Krishnan, and S. Krishnan, "Atomic-scale defects involved in the negative-bias temperature instability," *IEEE Trans. Device Mater. Rel.*, vol. 7, no. 4, pp. 540–557, Dec. 2007.

[16] J. F. Zhang, C. Z. Zhao, A. H. Chen, G. Groeseneken, and R. Degraeve, "Hole traps in silicon dioxides. Part I. Properties," *IEEE Trans. Electron Devices*, vol. 51, no. 8, pp. 1267–1273, Aug. 2004.

[17] C. Shen *et al.*, "Negative U traps in HfO_2 gate dielectrics and frequency dependence of dynamic BTI in MOSFETs," in *IEEE Int. Electron Devices Meeting (IEDM) Tech. Dig.*, Dec. 2004, pp. 733–736.

[18] D. S. Ang, S. Wang, G. A. Du, and Y. Z. Hu, "A consistent deep-level hole trapping model for negative bias temperature instability," *IEEE Trans. Device Mater. Rel.*, vol. 8, no. 1, pp. 22–34, Mar. 2008.

[19] D. Veksler and G. Bersuker, "Gate dielectric degradation: Pre-existing vs. generated defects," *J. Appl. Phys.*, vol. 115, no. 3, pp. 034517-1–034517-11, Jan. 2014.

[20] A. Stesmans, "Dissociation kinetics of hydrogen-passivated P_b defects at the (111)Si/SiO_2 interface," *Phys. Rev. B*, vol. 61, no. 12, pp. 8393–8403, 2000.

[21] A. Haggag, W. McMahon, K. Hess, K. Cheng, J. Lee, and J. Lyding, "High-performance chip reliability from short-time-tests-statistical models for optical interconnect and HCI/TDDB/NBTI deep-submicron transistor failures," in *Proc. 39th Annu. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr./May 2001, pp. 271–279.

[22] M. Houssa, "Modelling negative bias temperature instabilities in advanced p-MOSFETs," *Microelectron. Rel.*, vol. 45, no. 1, pp. 3–12, 2005.

[23] T. Grasser *et al.*, "Analytic modeling of the bias temperature instability using capture/emission time maps," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2011, pp. 27.4.1–27.4.4.

[24] T. Grasser, H. Reisinger, P.-J. Wagner, F. Schanovsky, W. Goes, and B. Kaczer, "The time dependent defect spectroscopy (TDDS) for the characterization of the bias temperature instability," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, May 2010, pp. 16–25.

[25] N. Goel, K. Joshi, S. Mukhopadhyay, N. Nanaware, and S. Mahapatra, "A comprehensive modeling framework for gate stack process dependence of DC and AC NBTI in SiON and HKMG p-MOSFETs," *Microelectron. Rel.*, vol. 54, no. 3, pp. 491–519, 2014.

[26] B. Kaczer *et al.*, "NBTI from the perspective of defect states with widely distributed time scales," in *Proc. Int. Rel. Phys. Symp. (IRPS)*, Apr. 2009, pp. 55–60.

[27] S. Desai *et al.*, "A comprehensive AC/DC NBTI model: Stress, recovery, frequency, duty cycle and process dependence," in *Proc. Int. Rel. Phys. Symp. (IRPS)*, Apr. 2013, pp. XT.2.1–XT.2.11.

[28] M. Toledano-Luque *et al.*, "Response of a single trap to AC negative bias temperature stress," in *Proc. Int. Rel. Phys. Symp. (IRPS)*, Apr. 2011, pp. 364–371.

[29] T. Grasser *et al.*, "Advanced characterization of oxide traps: The dynamic time-dependent defect spectroscopy," in *Proc. Int. Rel. Phys. Symp. (IRPS)*, Apr. 2013, pp. 2D.2.1–2D.2.7.

[30] A. Asenov, R. Balasubramaniam, A. R. Brown, and J. H. Davies, "RTS amplitudes in decananometer MOSFETs: 3-D simulation study," *IEEE Trans. Electron Devices*, vol. 50, no. 3, pp. 839–845, Mar. 2003.

[31] B. Kaczer *et al.*, "Origin of NBTI variability in deeply scaled pFETs," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, May 2010, pp. 26–32.

[32] J. Franco *et al.*, "Impact of single charged gate oxide defects on the performance and scaling of nanoscaled FETs," in *Proc. Int. Rel. Phys. Symp. (IRPS)*, Apr. 2012, pp. 5A.4.1–5A.4.6.

[33] T. Grasser, "Stochastic charge trapping in oxides: From random telegraph noise to bias temperature instabilities," *Microelectron. Rel.*, vol. 52, no. 1, pp. 39–70, 2012.

[34] T. Naphade, K. Roy, and S. Mahapatra, "A novel physics-based variable NBTI simulation framework from small area devices to 6T-SRAM," in *Proc. Int. Electron Devices Meeting (IEDM)*, Dec. 2013, pp. 33.6.1–33.6.4.

[35] T. Grasser *et al.*, "Switching oxide traps as the missing link between negative bias temperature instability and random telegraph noise," in *Proc. Int. Electron Devices Meeting (IEDM)*, Dec. 2009, pp. 1–4.

[36] A. Stesmans, "Passivation of P_{b0} and P_{b1} interface defects in thermal (100) Si/SiO_2 with molecular hydrogen," *Appl. Phys. Lett.*, vol. 68, no. 15, pp. 2076–2078, 1996.

[37] T. Grasser, H. Reisinger, K. Rott, M. Toledano-Luque, and B. Kaczer, "On the microscopic origin of the frequency dependence of hole trapping in pMOSFETs," in *Proc. Int. Electron Devices Meeting (IEDM)*, Dec. 2012, pp. 19.6.1–19.6.4.

[38] S. Chakravarthi, A. T. Krishnan, V. Reddy, C. F. Machala, and S. Krishnan, "A comprehensive framework for predictive modeling of negative bias temperature instability," in *Proc. Int. Rel. Phys. Symp. (IRPS)*, Apr. 2004, pp. 273–282.

[39] M. A. Alam, H. Kufluoglu, D. Varghese, and S. Mahapatra, "A comprehensive model for pMOS NBTI degradation: Recent progress," *Microelectron. Rel.*, vol. 47, no. 6, pp. 853–862, 2007.

[40] S. Choi, Y. J. Park, C.-K. Baek, and S. Park, "An improved 3D Monte Carlo simulation of reaction diffusion model for accurate prediction on the NBTI stress/relaxation," in *Proc. Simul. Semicond. Process. Devices*, Sep. 2012, pp. 185–188.

[41] F. Schanovsky and T. Grasser, "On the microscopic limit of the modified reaction-diffusion model for the negative bias temperature instability," in *Proc. Int. Rel. Phys. Symp. (IRPS)*, Apr. 2012, pp. XT.10.1–XT.10.6.

[42] B. Kaczer, V. Arkhipov, R. Degraeve, N. Collaert, G. Groeseneken, and M. Goodwin, "Disorder-controlled-kinetics model for negative bias temperature instability and its experimental verification," in *Proc. Int. Rel. Phys. Symp. (IRPS)*, Apr. 2005, pp. 381–387.

[43] S. Zafar, "Statistical mechanics based model for negative bias temperature instability induced degradation," *J. Appl. Phys.*, vol. 97, no. 10, pp. 103709-1–103709-9, May 2005.

[44] S. Mahapatra, V. D. Maheta, A. E. Islam, and M. A. Alam, "Isolation of NBTI stress generated interface trap and hole-trapping components in PNO p-MOSFETs," *IEEE Trans. Electron Devices*, vol. 56, no. 2, pp. 236–242, Feb. 2009.

[45] T. Naphade, N. Goel, P. R. Nair, and S. Mahapatra, "Investigation of stochastic implementation of reaction diffusion (RD) models for NBTI related interface trap generation," in *Proc. Int. Rel. Phys. Symp. (IRPS)*, Apr. 2013, pp. XT.5.1–XT.5.11.

[46] H. Reisinger *et al.*, "The effect of recovery on NBTI characterization of thick non-nitrided oxides," in *Proc. Int. Integr. Rel. Workshop*, Oct. 2008, pp. 1–6.

[47] S. Mahapatra, P. B. Kumar, and M. A. Alam, "Investigation and modeling of interface and bulk trap generation during negative bias temperature instability of p-MOSFETs," *IEEE Trans. Electron Devices*, vol. 51, no. 9, pp. 1371–1379, Sep. 2004.

[48] S. Mahapatra *et al.*, "A critical re-evaluation of the usefulness of R-D framework in predicting NBTI stress and recovery," in *Proc. Int. Rel. Phys. Symp. (IRPS)*, Apr. 2011, pp. 6A.3.1–6A.3.10.

[49] T. Grasser *et al.*, "Hydrogen-related volatile defects as the possible cause for the recoverable component of NBTI," in *Proc. Int. Electron Devices Meeting (IEDM)*, Dec. 2013, pp. 15.5.1–15.5.4.

[50] J. Franco *et al.*, "NBTI reliability of SiGe and Ge channel pMOSFETs with SiO_2 HfO_2 dielectric stack," *IEEE Trans. Device Mater. Rel.*, vol. 13, no. 4, pp. 497–506, Dec. 2013.

[51] T. Grasser *et al.*, "The 'permanent' component of NBTI: Composition and annealing," in *Proc. Int. Rel. Phys. Symp. (IRPS)*, Apr. 2011, pp. 605–613.

Authors' photographs and biographies not available at the time of publication.