Critical Modeling Issues in Negative Bias Temperature Instability

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

* CDL for TCAD at the Institute for Microelectronics, TU Wien, Wien, Austria
° IMEC, Leuven, Belgium

Both the physical mechanisms as well as the modeling of negative bias temperature instability (NBTI) have attracted growing attention during the last years. While the reaction-diffusion theory had been the dominant explanation for a relatively long period, a growing number of authors have recently voiced their doubts regarding its validity. We give a brief review of suggested models and highlight their strengths and, more importantly, their weaknesses. We take care not to get lost in the intricacies of the various models by only qualitatively discussing their features. As will be shown, this is more than sufficient to demonstrate considerable shortcomings. Finally, we summarize our latest modeling attempts which try to overcome the observed modeling contradictions and show comparisons to experimental data.

Introduction

Negative bias temperature instability (NBTI) has been known for more than forty years but has attracted growing attention recently. After a period when the reaction-diffusion (RD) theory (1, 2) reigned more or less undisputedly as the dominant explanation, a growing number of authors (3–9) have recently voiced their doubts regarding its validity. In particular, whether NBTI is due to interface states and/or oxide charges is amongst the most controversial issues at the time. This recent controversy has also been fueled by the introduction of new fast measurement techniques, which are capable of monitoring degradation and recovery in the microseconds regime.

In order to estimate device lifetimes, constant bias and temperature stress is conventionally employed. The stress bias is then interrupted at predefined times to determine the degradation using a measure/stress/measure kind of technique (MSM)(10). Alternatively, the degradation is monitored without the interruption of stress by using on-the-fly (OTF) techniques (11). Conventionally, the shift in the threshold-voltage is taken as a measure for degradation, although other critical device parameters degrade as well, for instance the on-current and the mobility. While OTF techniques avoid recovery, the conversion of the recorded drain-current degradation to a threshold-voltage shift is highly controversial (12–14). In addition, changes in the charge-pumping currents are often recorded in order to gain information on the relative importance of interface versus oxide charges.

Unfortunately, these 'constant stress condition' setups have also traditionally been used for the development of models. Recently it has been recognized that these models fail to explain a number of features visible only under dynamic boundary conditions (recovery induced by changes in the bias conditions, duty-factor dependent stress, etc.).

Examples are the ubiquitous log-like recovery observed to cover at least twelve decades in time (5, 8), the strong bias sensitivity particularly to positive biases (15), a marked duty-factor dependence (15–17), the presence of possibly two contributions (e.g., oxide and interface charges) (4, 5, 18), and the initial log-like vs. the long term power-law-like behavior (18–20). Consequently, we believe that a good understanding of the phenomenon can only be developed by studying the degradation response to dynamic bias conditions.

Some critical issues which are only poorly understood at the time are

- Is NBTI due to interface and/or oxide charges?
- Why does recovery take much longer than degradation?
- Why does recovery follow a ubiquitous logarithmic behavior in various technologies, over a broad range of bias conditions and temperatures?
- Why does recovery strongly depend on gate bias, in particular positive bias conditions?
- Why does the degradation show such a marked duty-factor dependence?

A number of models have been suggested to answer at least some of the questions raised above. We have implemented these models in our numerical device and circuit simulator MINIMOS-NT (21) to fully understand their behavior without having to resort to simplified analytical approximations. We have then benchmarked these models against real data to see if they live up to their promises. In this review we will not ponder upon the exact definition and details of these models nor on the differential equation sets that describe them, which can be found in the referenced literature. Rather, we will qualitatively describe the basic ideas of the model using energy diagram schematics which eventually determine the model behavior, irrespective of minor details and choices of parameters. Nonetheless, numerical simulation results will be given which undisputedly demonstrate that most models are insufficient. Possible alternative modeling attempts will be suggested and discussed.

Universal Relaxation

Empirical analysis of a large set of experimental data revealed that NBTI recovery follows a universal pattern (6, 15, 22, 23): when the relaxation data $\Delta V_{\mathrm{th}}(t_{\mathrm{s}}, t_{\mathrm{r}})$ as a function of the relaxation time t_{r} taken after a stress time t_{s} are normalized to the recovery-free data ($t_{\mathrm{r}} = 0$) and plotted versus the ratio of the relaxation to the stress time ($\xi = t_{\mathrm{r}}/t_{\mathrm{s}}$), all data fall on a single universal curve. As an empirical expression for this universal relaxation law

$$r(\xi) = \frac{\Delta V_{\mathrm{th}}(t_{\mathrm{s}}, t_{\mathrm{r}})}{\Delta V_{\mathrm{th}}(t_{\mathrm{s}}, 0)} = \frac{1}{1 + B\xi^{\beta}} \qquad [1]$$

has been suggested (6) with B and β as fit parameters. Later (15, 23), a permanent component has been added to the model,

$$\Delta V_t(t_s, t_r) = R(t_s) r(t_r/t_s) + P(t_s) \ . \qquad [2]$$

Unfortunately, the model is purely empirical and therefore does not provide much insight into the underlying microscopic processes, except for the fact that they must be able

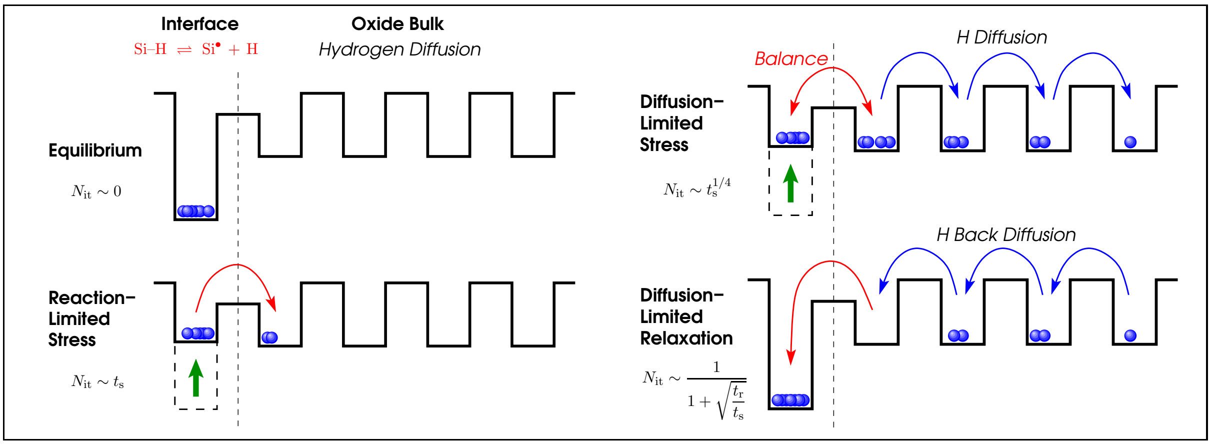
Interface
$Si–H \rightleftharpoons Si^{\bullet} + H$
Oxide Bulk
Hydrogen Diffusion
Equilibrium
$N_{\mathrm{it}} \sim 0$
Reaction–Limited Stress
$N_{\mathrm{it}} \sim t_{\mathrm{s}}$
Balance
H Diffusion
Diffusion–Limited Stress
$N_{\mathrm{it}} \sim t_{\mathrm{s}}^{1/4}$
H Back Diffusion
Diffusion–Limited Relaxation
$N_{\mathrm{it}} \sim \frac{1}{1+\sqrt{\frac{t_{\mathrm{r}}}{t_{\mathrm{s}}}}}$

Figure 1: Schematic view of the atomic hydrogen reaction-diffusion model during stress and relaxation.

to accommodate a large number of different timescales. It may nevertheless serve as a touchstone for more detailed physics-based models. In particular, as will be shown in the following, all extensions of the reaction-diffusion theory can be dismissed based on this observation as they all introduce non-universal humps, that is, characteristic features which occur at different normalized times, into the recovery characteristics (6, 24, 25).

Reaction-Diffusion Theory Based Models

Until recently, the reaction-diffusion theory, or variants thereof, have been used almost exclusively for the explanation of NBTI. The basic idea behind this theory is summarized as follows: initially, all existing interface states are passivated by hydrogen. Upon application of stress, the bonds are rapidly broken via a thermally and field-activated process and the released hydrogen starts to diffuse away, either as atomic hydrogen or as H_2. The diffusion of hydrogen is assumed to be slow and to control the overall dynamics. The assumption of classical diffusion appears to be in contradiction to reports who have demonstrated that hydrogen is a highly reactive species that readily interacts with the medium it diffuses in (26, 27). As such, it is assumed that defect creation only occurs at the interface in the form of dangling bonds.

In terms of energy levels this implies that the single-valued diffusion barrier for hydrogen is considerably larger than the binding energy of hydrogen at the interface state during stress. Since diffusion of neutral hydrogen (H^0 or H_2) is commonly assumed, the diffusion barriers do not change with field. Due to this considerable difference in barrier heights, the forward and backward reactions at the interface are in quasi-equilibrium and the density of depassivated interface states and the interfacial hydrogen concentration is dictated by the mass-action law. Consequently, the mechanism is controlled by the concentration of hydrogen at the interface, which in turn is assumed to follow a simple standard diffusion equation (1).

The simplest variant of such a model assumes diffusion of atomic hydrogen and is schematically depicted in Fig. 1. During the initial phase, which is very short and has never been experimentally observed, the interface reaction comes into quasi-equilibrium. During this phase, the degradation is directly proportional to the stress time t_{s}. Once

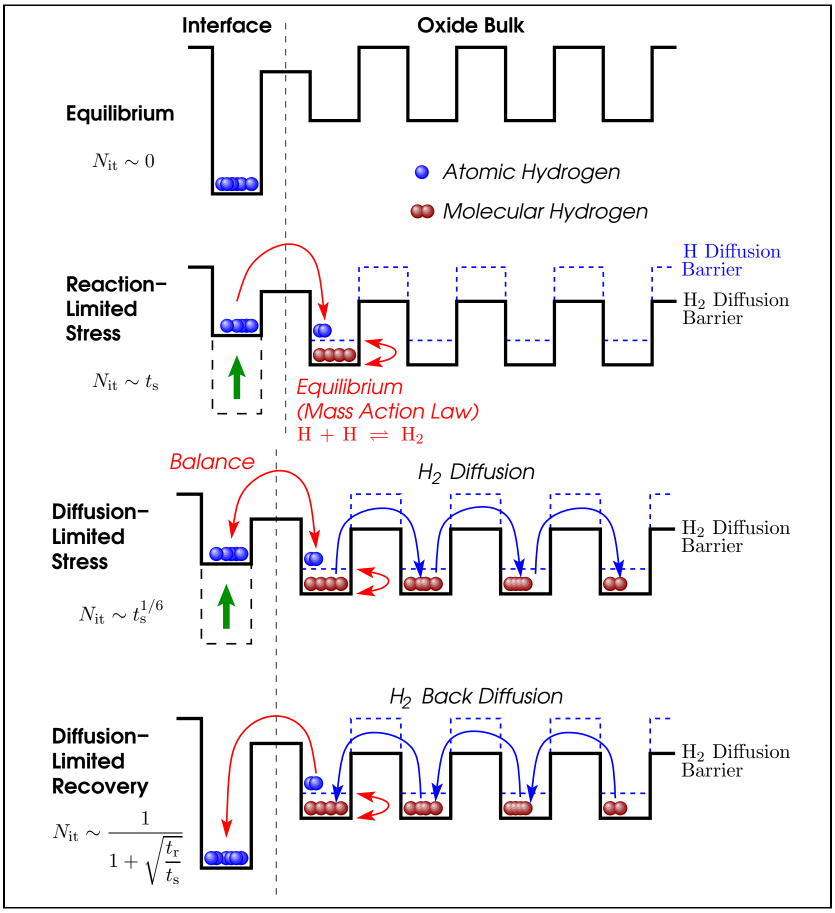

Figure 2: Schematic view of the molecular hydrogen reaction-diffusion model during stress and relaxation.

quasi-equilibrium has been reached, the degradation is controlled by the diffusion of hydrogen and a power-law dependence of the form $t_{\rm s}^{1/4}$ is obtained. As soon as the stress is turned off, quasi-equilibrium at the interface requires that all interfacial hydrogen passivates an interface state, and thus the hydrogen concentration at the interface approaches zero very quickly (at timescales well below any measurement resolution with the parameters conventionally used). Then, the refilling of this interfacial hydrogen layer by backdiffusion controls the overall recovery. During this diffusion-limited recovery phase, hydrogen also keeps diffusing away from the interface (28), making the recovery rate *independent* of the diffusion constant, which effectively cancels out (29). Furthermore, due to the very large backward rate during recovery, the actual values of the rates entering the mass-action law become basically irrelevant.

The atomic hydrogen version of the RD theory predicts the degradation to follow $t_{\rm s}^{1/4}$, which is in decent agreement with measurement data acquired using larger delay times. Faster measurements produce data with considerably smaller power-law exponents, and the discovery of this effect triggered a refinement of the RD theory (30). In the refined version, atomic hydrogen is assumed to dimerize instantly, thereby forming H_2 with the overall degradation being now controlled by the diffusion of H_2, see Fig. 2. Due to the change in the mass-action law, where now the square-root of the H_2 concentration

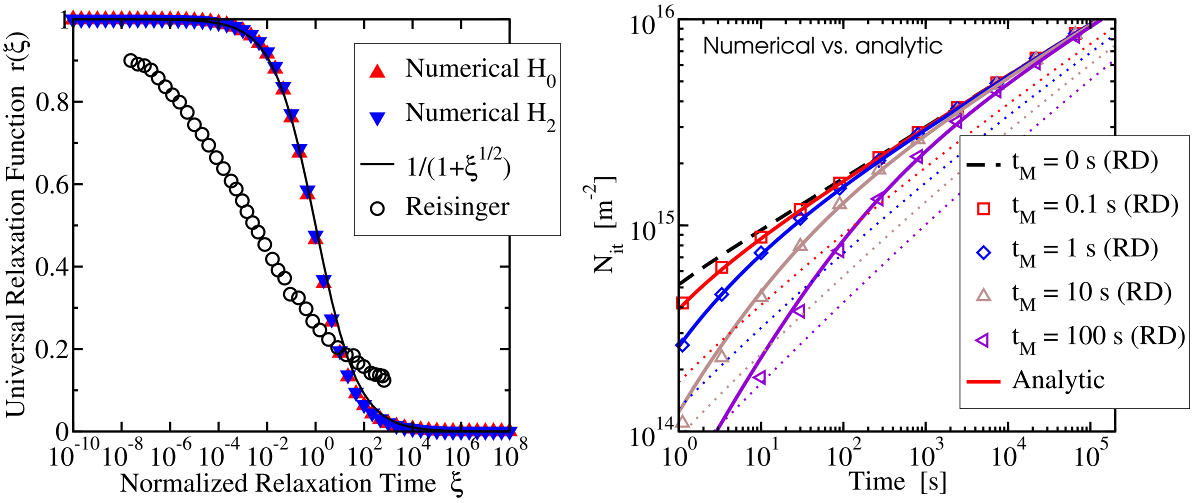

Figure 3: Left: Recovery simulated with the H and H_2 RD models in comparison to a typical experimental recovery trace take from Reisinger *et al.* (5). The simulated recovery occurs too late and then too fast. **Right**: The impact of a measurement delay on the simulated RD degradation vanishes after $t_s = 10^4$ s, in contrast to experimental data where the delay also has a significant impact after 10^4 s, as schematically indicated by the dotted lines.

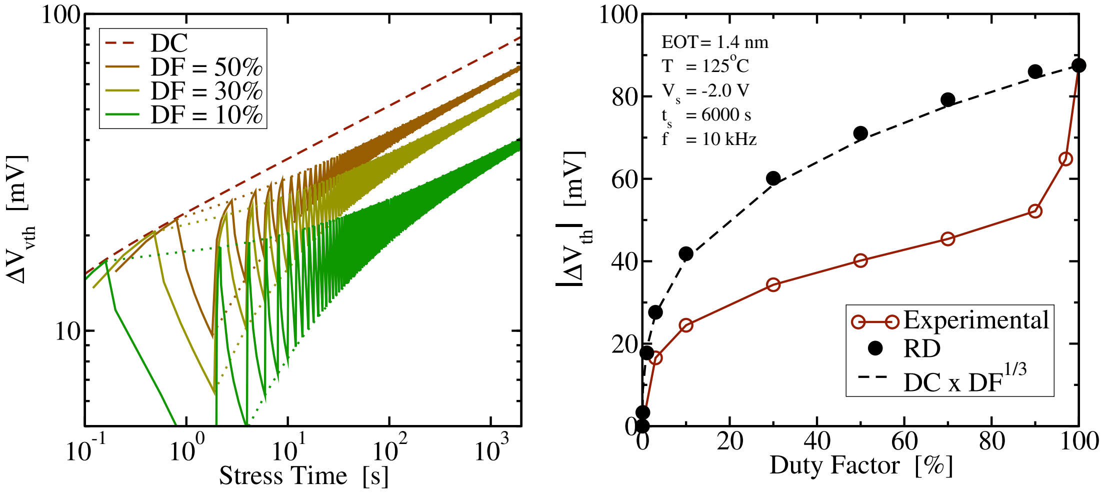

Figure 4: As a direct consequence of the failure to correctly describe recovery, the duty-factor (DF) dependence of NBTI is also wrongly predicted by RD models. **Left**: Simulated impact of DF stress in the time-domain. **Right**: Simulated degradation level after 6000 s stress as a function of the DF. Experimental data show a very strong sensitivity around DF=100% (15, 16), which is completely absent in the RD simulation. In contrast to the data, the prediction of the RD model is proportional to $DF^{1/3}$.

replaces the concentration of H, a power-law exponent of 1/6 is obtained, which is closer to faster experimental data. The recovery, on the other hand, is *not affected* by this model refinement since the actual values of the rates only impact the very early recovery phase, which occurs much earlier than the diffusion controlled recovery can set in and is only responsible for a small fraction of the total recovery.

During recovery, both the H and the H_2 version are universal in the sense of [1] and well approximated by (6)

$$r(\xi) \approx \frac{1}{1+\xi^{1/2}} \qquad [3]$$

As shown in Fig. 3, expression [3] predicts relaxation to occur within 4 decades in time, in stark contrast to experimental data which cover at least 12 decades. Furthermore, the predicted recovery does not depend on voltage and temperature, nor does it depend on *any* parameters of the model. As such, it cannot explain any of the experimental recovery data which show a marked voltage dependence, in particular at positive bias, a temperature dependence, and also considerable dependence on the details of the fabrication process. Additional manifestations of basically the same problem are shown in Fig. 4, where the RD model is used to predict the impact of duty-factor (DF) dependent stress. During DF stress the stress is periodically interrupted with a switch back to the threshold voltage to emulate the on-time experienced by a transistor in a real circuit. Again, poor agreement with experiment is obtained.

Numerous attempts have been made to salvage the model by adding further refinements. None of them solves the basic problem of the model as will be demonstrated in the following. In fact, most refinements even break the universality and are in that sense even worse than simple RD models.

Two-Region RD Model: Diffusion in the Polysilicon Gate

With the advent of modern technology nodes it was realized that it is impossible to argue in favor of hydrogen diffusion in oxide layers only a few nanometers thick and it was suggested that the diffusion also occurs in the polysilicon gate (31). By assuming different diffusivities in SiO_2 and the polysilicon gate, this was considered a chance to introduce faster recovery than that normally predicted by RD theory and to explain fast and slow components. However, as we will demonstrate, none of this appears to be the case.

The simplest refinement of the RD model is built around the assumption of two different diffusivities in the oxide and the polysilicon gate. Such a two-region model based on H_2 diffusion is schematically shown in Fig. 5. During the initial stress, the oxide is rapidly filled with hydrogen. During that phase, the model behaves just like the standard H_2 RD model. As soon as the diffusion front reaches the polysilicon interface, H_2 starts piling up, thereby slowing down the degradation by lowering the concentration gradient. Eventually, H_2 starts spilling over the interface, thereby creating a new diffusion front. As soon as the H_2 concentration in the polysilicon dominates over the H_2 concentration in the oxide, the model again behaves like the standard H_2 RD model (31), see Fig. 6. Such an intermediary kink introduced by the transition of these two regimes during stress, however, has never been experimentally observed.

During recovery, the 'faster' H_2 inside the oxide results in accelerated recovery which sets in earlier than predicted by [3]. In fact, the functional form of this initial recovery still follows [3], when the stress time t_s is replaced by the time the initial degradation started to pile up and temporarily saturate. Depending on the ratio of H_2 stored in the oxide and the polysilicon regions, the recovery will either follow the traces of the saturated

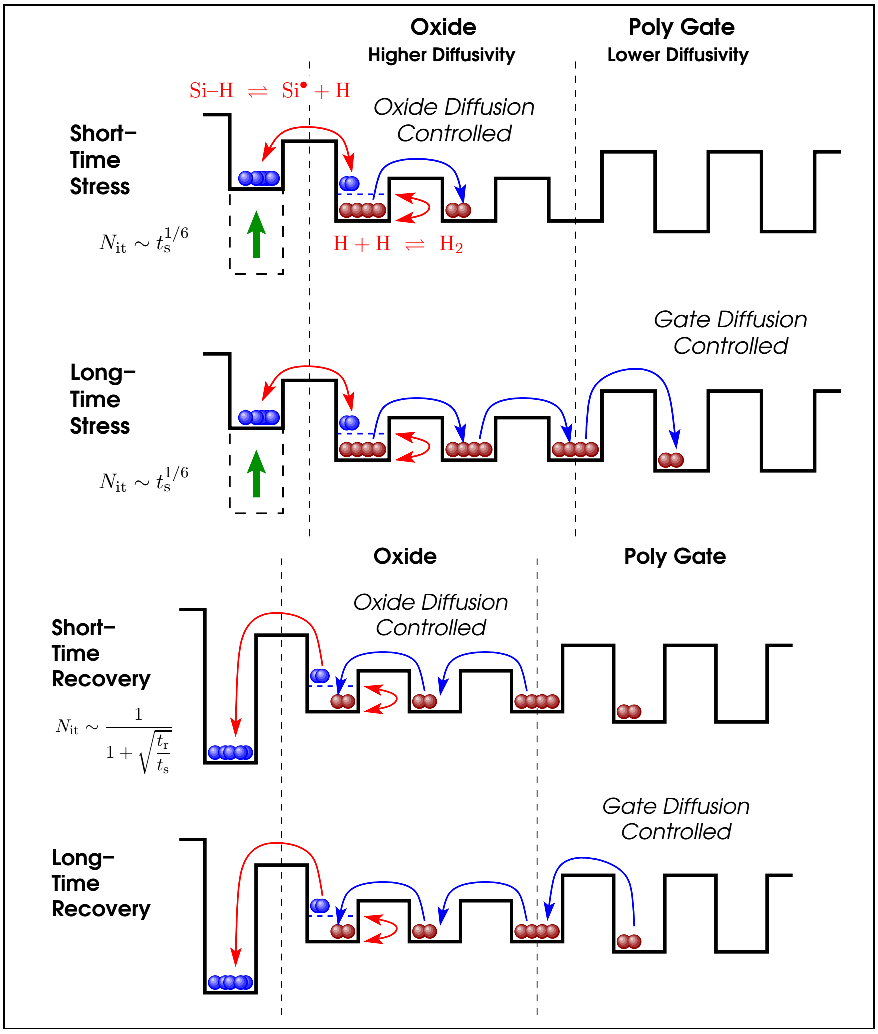

Figure 5: Schematic view of the two-region RD model which assumes a different diffusivity in SiO_2 and the polysilicon gate, again during stress and recovery. Due to the higher diffusivity in the SiO_2 layer, return of H stored in this layer is faster, thereby (marginally) accelerating recovery.

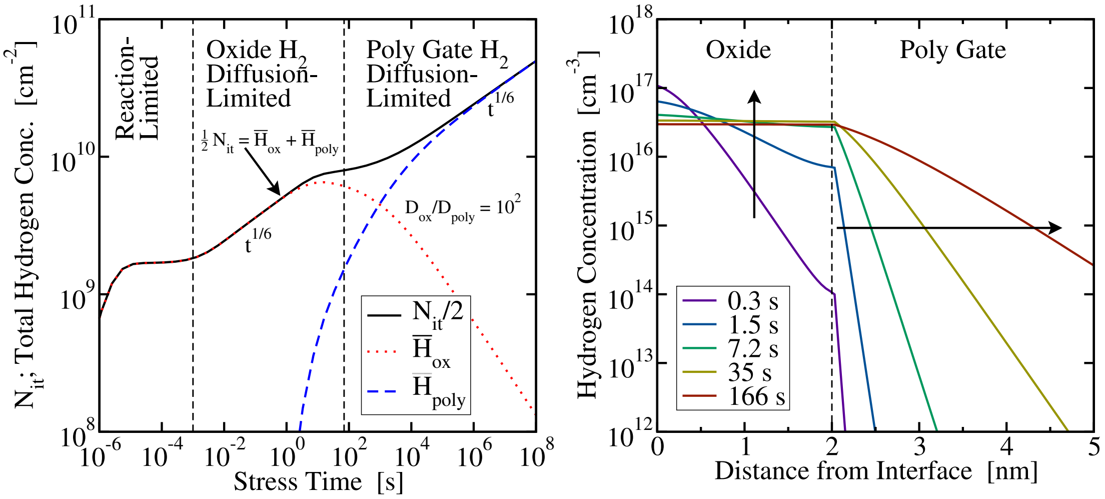

Figure 6: Simulated behavior of the two-region RD model during stress. Initially, H_2 is built-up inside the oxide and once filled, hydrogen diffusion in the polysilicon begins to dominate the degradation. This assumption results in a kink during degradation (at about 100 s in the above example), not yet experimentally observed. (The slowdown will turn into a speed-up for the reversed case where the diffusivity in the polysilicon is larger than in the oxide.)

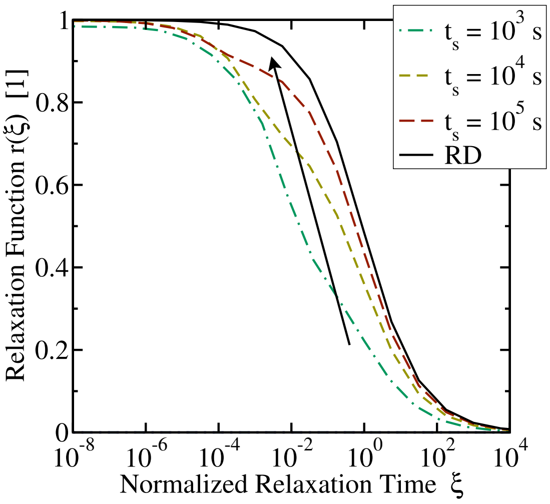

Figure 7: Simulated recovery behavior of the two-region RD model. Introduction of the two layers only marginally increases the range of time-scales and spoils the universality.

population, or, for longer stress, behave just like the standard H_2 RD recovery as soon as the H_2 concentration in the oxide can be neglected over the concentration inside the polysilicon. Since the amount of H_2 in the oxide is basically saturated while only the population in the polysilicon grows, the recovery will not be universal, in contrast to experimental data, cf. Fig. 7.

Two-Interface RD Model: Impact of the Oxide/Polysilicon Interface

In another attempt to improve the original RD model, the two-region model was refined based on the observed increase of stress-induce leakage current after NBTI stress, which suggests trap generation also at the oxide/polysilicon interface (32). The model assumes that atomic hydrogen is released from the silicon/oxide interface, quickly diffuses through the oxide and plucks off a hydrogen atom from a previously passivated interface state. The thereby formed H_2 then diffuses into the polysilicon, see Fig. 8. As with the two-region model, this model was suspected to explain the fast recovery seen in experiments. During stress, after an initial plateau, this model also behaves like the standard H_2 model, see Fig. 9. During recovery, the hydrogen stored in the oxide layer causes a fast initial recovery, see Fig. 10. Contrary to experiment, though, the relative importance of this fast initial recovery diminishes with increasing stress time. This is because the generated interface states in the RD model are exactly equal to the amount of released hydrogen and since the hydrogen concentration in the oxide reaches a quasi-equilibrium value, the amount of interface states that can be passivated using this hydrogen decreases with increasing stress. Thus, non-universal humps are introduced.

RD Model with Explicit Dimerization

Another refinement proposed to improve the model prediction during the initial stress phase is based on the idea that dimerization does not occur instantly and consequently both H^0 and H_2 can exist and diffuse in parallel (33). This is schematically visualized in Fig. 11. Since the long-term power-law exponent given by the H_2 model should be maintained, the parameters of the refined model have to be chosen in a way to guarantee quasi-equilibrium of the dimerization rate equation at longer stress times. Thereby, only the initial degradation phase is modified, leading to a short region with a power-law with

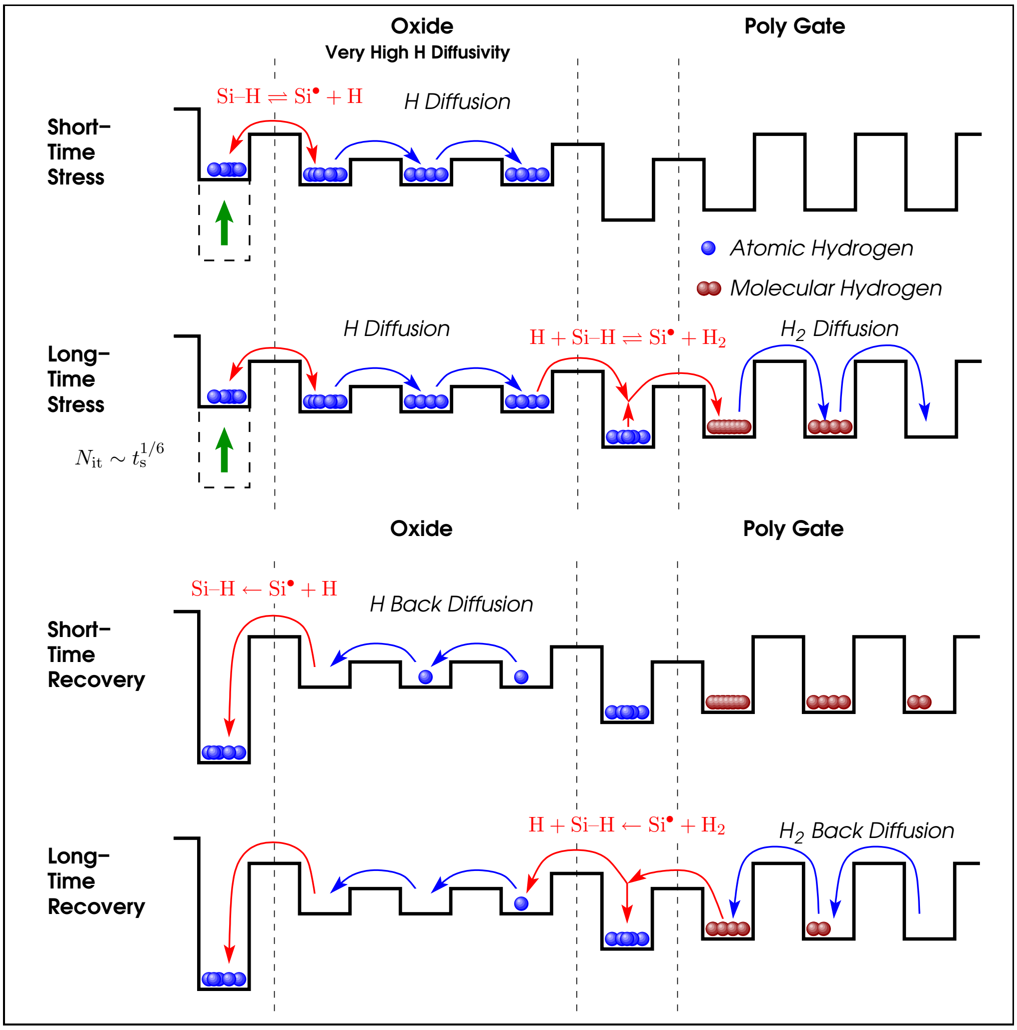

Figure 8: Schematic view of the two-interface RD model during stress and relaxation.

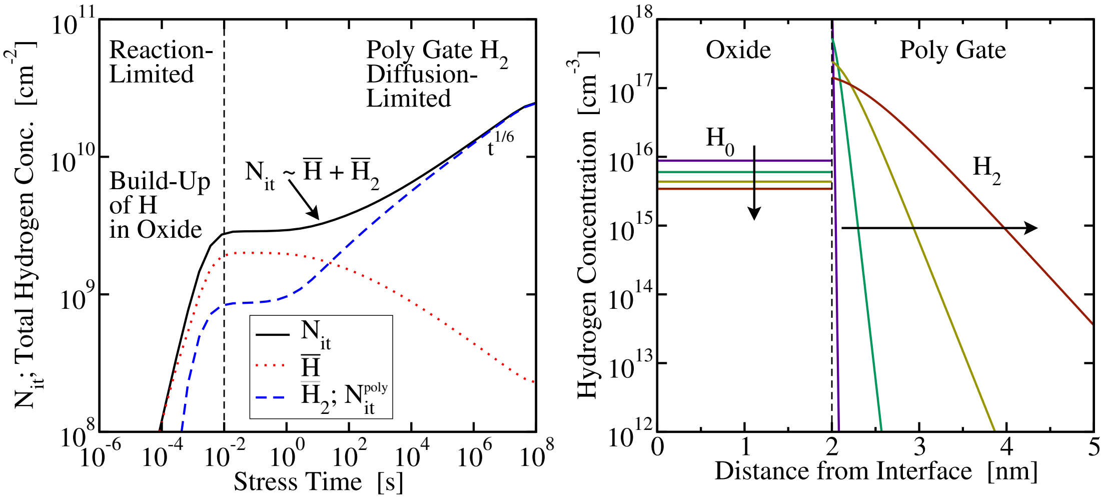

Figure 9: Simulated degradation characteristics with the two-interface RD model. Initially, atomic hydrogen is built-up in the oxide which eventually swaps into the polysilicon as H_2 after having plucked a passivating H from dangling bonds at the oxide/polysilicon interface. The hydrogen concentration in the oxide reaches a quasi-saturation value and the model behaves basically like the standard H_2 model. The right figure shows the hydrogen profile in the oxide and the polysilicon at for different times after the build-up of quasi-saturation ($t_s > 100\,\mathrm{s}$), where the arrows indicate the temporal evolution.

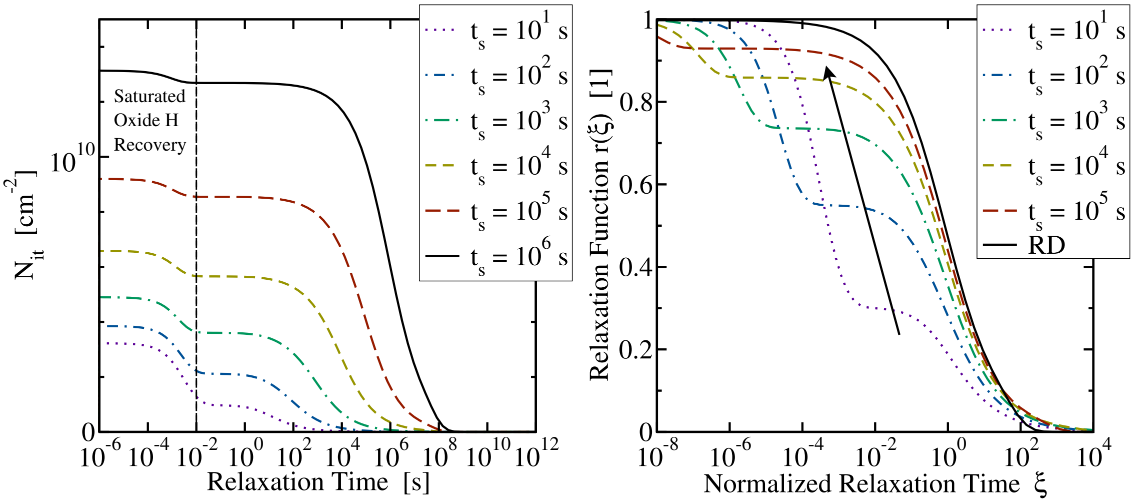

Figure 10: Simulated recovery using the two-interface RD model. The fractional importance of the fast recovery diminishes with stress time and introduces non-universal humps.

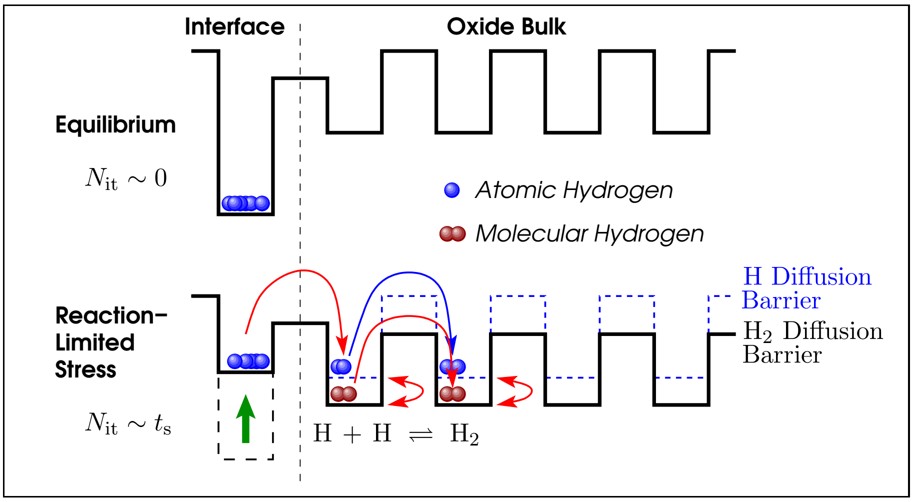

Figure 11: Schematic view of the H^0/H_2 reaction-diffusion model during stress where diffusion of both atomic and molecular hydrogen is considered.

exponent 1/3. This was claimed to be in agreement with experimental data (33). It has also been speculated that the existence of both H^0 and H_2 populations may allow a description of fast recovery. Quite to the contrary, however, a rigorous simulation reveals that there is *no impact on recovery whatsoever*, see Fig. 12, and the model predicts the same (wrong) recovery as the standard H^0 and H_2 models.

RD Model with Interface State Occupancy Dynamics

In an attempt to explain the ubiquitous log-like recovery characteristics observed following NBTI stress, it was suggested that this may be a measurement artifact as the occupancy of the interface states cannot follow the quick changes of the gate voltage (12): Right after stress, all interface states are positively charged (the lone electron of the dangling bond is missing) and as during recovery the Fermi-level is moved above the silicon valence band, electrons are captured (the hole is emitted), which indeed results

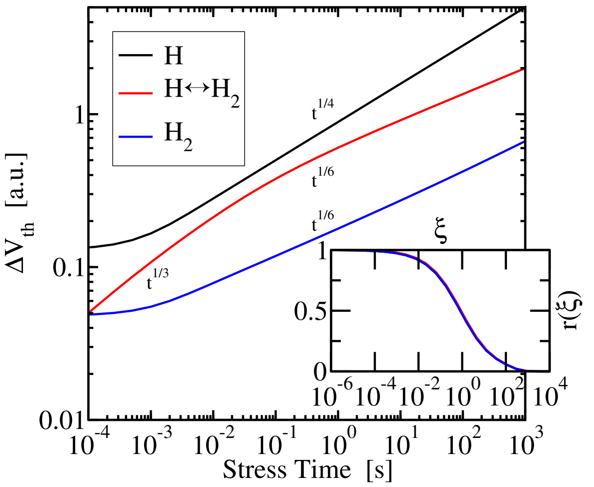

Figure 12: Simulated degradation using the H^0/H_2 RD model. With the conventional choice of parameters the model stays between the H^0 and the H_2 RD models, with the initial behavior given by $t_s^{1/3}$ and the long term behavior close to the H_2 model, $t_s^{1/6}$. As the model remains in the boundaries set by the H^0 and the H_2 models, no impact on recovery is obtained.

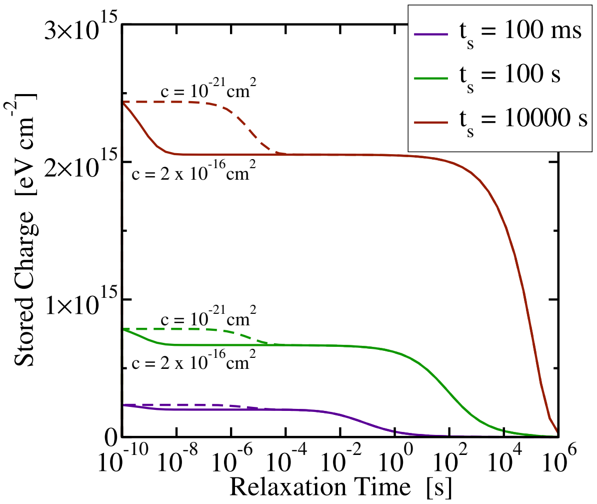

Figure 13: Simulated recovery when the occupancy of the interface states is given by SRH statistics. For conventionally used capture cross sections ($2 \times 10^{-16}\,\text{cm}^2$) the response of the interface states for a switch from stress to relaxation are in the nanosecond range (first step). Only excessively small values (e.g. $1 \times 10^{-21}\,\text{cm}^2$ (12)) result in measurable impact, at the expense of lost universality, though. Real recovery predicted by the RD model gives the second step.

in a $\log(t_\mathrm{r})$ component. Although such a charging transient is undeniable, the extensive literature available on interface state charging dynamics using Shockley-Read-Hall theory (SRH) clearly states that the transients due to electron capture are well within the nano-second range for the bias changes considered here and *do not cause any error in the measurements*, see Fig. 13. For example, the typical maximum time constant for interface states is in the millisecond regime for traps located at mid-gap (34). During a typical bias-switch in NBTI measurements, however, the Fermi-level is moved from below the valence band (stress) to the threshold level, which is not too far from the valence band either. Thus, the maximum time constant of these charging transients is small and also *independent* of stress time, again resulting in non-universal humps in relaxation transients.

RD Models with Dispersive Transport

It has long been understood that hydrogen motion through amorphous materials is a rather complex phenomenon. This is because hydrogen readily reacts with the surrounding lattice, thereby breaking and rearranging bonds. Furthermore, hydrogen can exist in various charge states, H^+, H^0, and H^-, depending on the local chemical potential. Various attempts have been made to describe hydrogen transport but due to its complexity empirical models are often used. Interestingly, hydrogen transport in SiO_2 has not been as well investigated as for instance in amorphous hydrogenated silicon or polysilicon and no universally accepted model exists for any of those material systems (27, 35). A common feature to modeling attempts, though, is the existence of a broad density of states for possible hydrogen capture sites. Unfortunately, this density of states may change

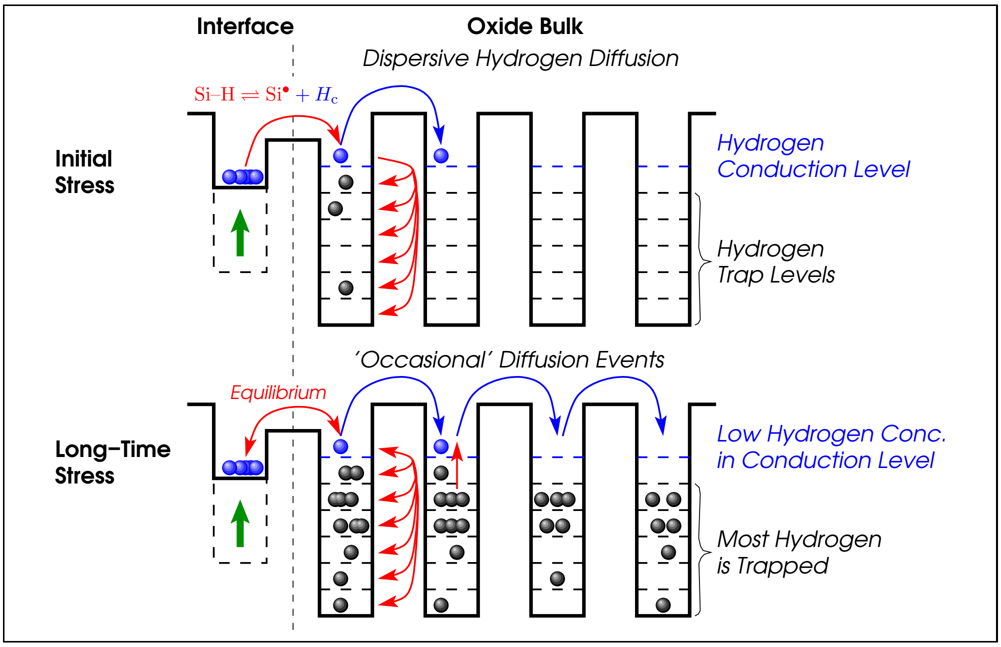

Figure 14: Schematic view of a RD model using the multiple-trapping dispersive transport model. Inside the oxide, hydrogen transport occurs in the hydrogen conduction band while most hydrogen resides on deeper levels. Thermal activation raises hydrogen to the transport level which is also used for the backward and forward reactions at the interface.

with hydrogen motion. Depending on the concentration of hydrogen, hydrogen diffusion may appear classical in the high concentration limit, or dispersive, for low concentrations (36–40).

Classical diffusion refers to the standard diffusion equation based on a single transport level as has been schematically used for instance in Figs. 1 and 2. This means that for each 'hop' of a hydrogen atom to a neighboring position, the same barrier height has to be surmounted, resulting in a uniform motion. For example, an initial concentration peak will result in the typical Gaussian broadening with time. In contrast, dispersive transport proceeds via a broad distribution of energy levels, which implies that each hop may face a vastly different energy barrier. Some of these hops will occur rapidly, leading to quick filling of shallow states. As time progresses, more and more deep states will be filled, resulting in a dramatic slow-down of the average hydrogen motion (41). Empirically, this can be approximately described by a time-dependent diffusivity $D(t) = D_0(\nu_0 t)^{\beta-1}$, with β being a temperature-dependent dispersion parameter, ν_0 an attempt frequency, and D_0 a microscopic diffusion constant (42).

More rigorous dispersive transport models were first applied to describe the movement of holes in amorphous materials (37) and H^+ after irradiation damage (40). While the first studies were based on the continuous time random walk (CTRW) theory developed by Scher and Montrol (37, 40), multiple-trapping (MT) models were proposed soon afterwards (38, 39, 41). Both models exhibit similar features (43–45) and their simplified versions were used to describe NBTI (46–50).

Based on the above given ideas, extended reaction-dispersive-diffusion models (RDD) have been proposed (10, 46–48). In these models it has been argued that transport of

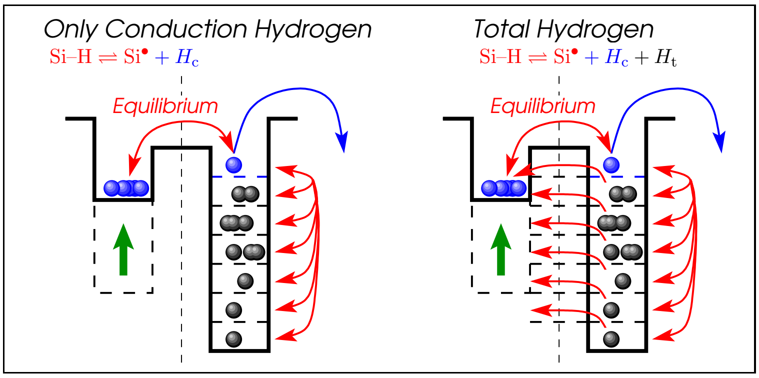

Figure 15: Comparison of the two boundary conditions used so far in dispersive RD models. **Left**: Models based on the multiple-trapping equations (see Fig. 14) assume that only hydrogen residing in the transport level interacts with the interface states. **Right**: In simplified models based on a time-dependent diffusivity it is implicitly assumed that *all* hydrogen, regardless of the depth of the trap level can move back to the interface at the same rate.

the hydrogen species inside the oxide is dispersive, consistent with hydrogen diffusion measurements (27, 35) and available models for irradiation damage (40, 51). Interestingly, in these models the power-law exponent depends on a temperature-dependent dispersion parameter. One feature common to published trap-controlled dispersive NBTI models is that they predict a reduction of the power-law exponent with increasing dispersion (10, 47, 48). However, in contrast to that it was observed that inclusion of traps into the standard RD model *increases* the exponent (30). In addition, a straight-forward application of the dispersive multiple-trapping transport model (38, 39, 41) also *increases* the exponent (52), in contradiction to these published reaction-dispersive-diffusion (RDD) models (10, 46–48).

A detailed analysis has revealed that the boundary condition at the Si/SiO_2 interface is the main reason for this discrepancy (50). In the variant where the traps are explicitly accounted for in the model (see Fig. 14), only hydrogen residing in the hydrogen conduction band (the transport state) is allowed to re-passivate an interface state. This slows down the reverse rate of the RD model as most hydrogen resides on the trap levels, while the forward rate is unaffected. Consequently, the degradation is accelerated and the power-law time exponent *increases*. In addition, the model predicts this power-law exponent to decrease with temperature, which is in contradiction to all currently available observations (4, 10, 53).

In contrast, simplified models do not normally differentiate between trapped and free hydrogen but only consider a total hydrogen concentration (49). This total hydrogen concentration is then used in the backward rate for the RD interface reaction. Microscopically this implies that all hydrogen regardless of its binding level can go back to the interface with the same probability, see Fig. 15. As a consequence, since the traps store a considerable amount of hydrogen directly at the interface, the backward rate is dramatically increased resulting in a decreased overall degradation. Furthermore, the power-law slope is predicted to increase linearly with temperature, consistent with experimental results obtained with delayed measurements (4, 10), but inconsistent with the delay-free measurement results of (18, 53).

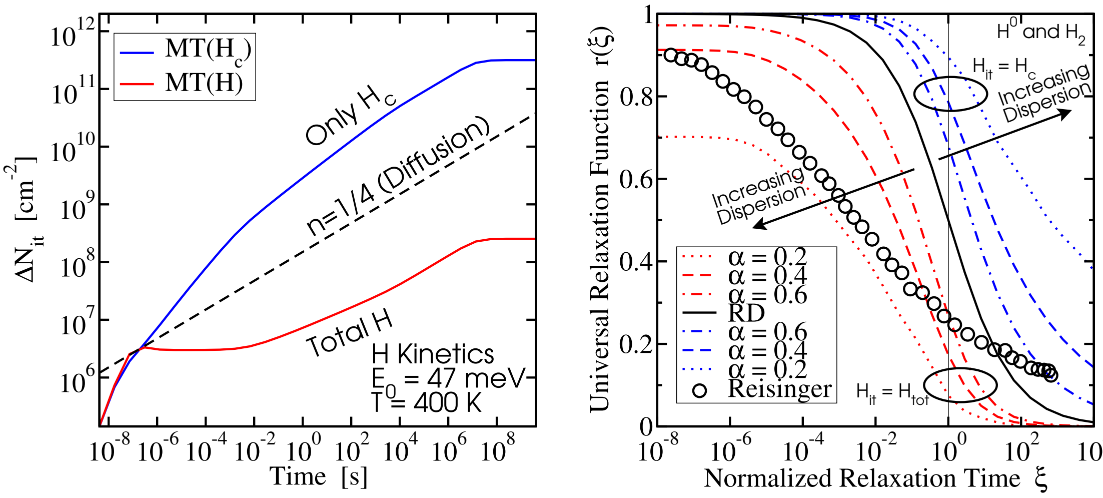

Figure 16: Simulated degradation and recovery using the two reaction-dispersive-diffusion (RDD) models. **Left**: Models allowing only the hydrogen in the conduction band to passivate an interface state predict an increase in the power-law slope during stress while models using the total amount of hydrogen for passivation show decreased degradation. **Right**: During recovery the trends are reversed: models based on the conduction band hydrogen show a retarded recovery as the hydrogen has to be first released from the trap level to the conduction band from where it can re-passivate an interface state. Models using the total hydrogen concentration show a rapid initial recovery from the hydrogen stored directly at the interface (time constants outside the visible are shown above) followed by a slower recovery controlled by dispersive back diffusion. Although the recovery can be slowed down and thus made to cover more decades, neither model is consistent with experimental data.

Reversed trends are observed during recovery (Fig. 16): in models that only allow hydrogen from the hydrogen conduction band to re-passivate an interface state, the recovery is delayed as the hydrogen from the deep traps has to be thermally activated to the hydrogen conduction band first before re-passivation becomes possible. In contrast, when a model does not differentiate between trapped and free hydrogen, a rapid initial recovery is observed which is then followed by a similar recovery controlled by dispersive back-diffusion. Interestingly, the recovery predicted by the standard RD model (the trap-free case) is like a water-shed which lies exactly between those two variants and cannot be crossed by either model. As such, total hydrogen models always recover too fast while conduction band hydrogen models can produce very long relaxation tails but lack any fast initial recovery. Consequently, none of these models is able to explain the experimental data.

General Conclusions on RD Models

We have shown that irrespective of the extensions applied to the RD model, the recovery behavior observed during measurement cannot be described with the published RD variants in their present form. The fact that some OTF measurements and fast MSM measurements give exponents of around $n = 0.15$, which is close to the value predicted by the H_2 based RD model ($n = 1/6$), should not let one arrive at the conclusion that the RD model is consequently reasonable. In particular, we think one has to be extremely cautious with a point of view that the RD model correctly covers the stress part while

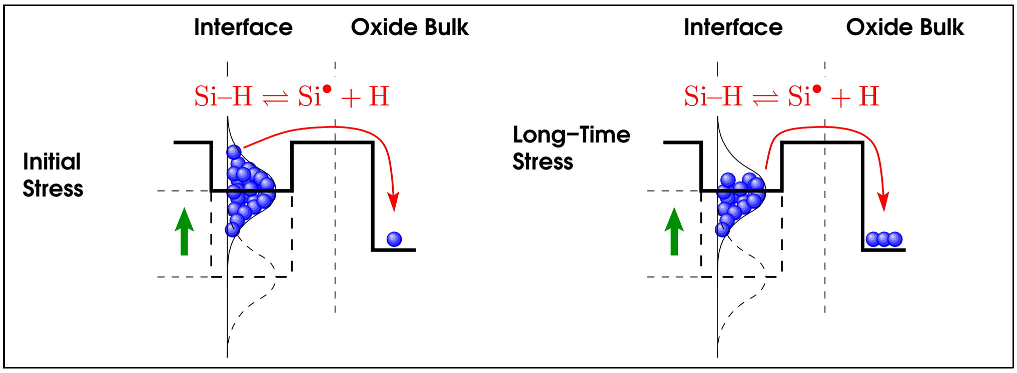

Figure 17: Schematic view of the dispersive bond breaking mechanism at the interface. Initially, weaker bonds are broken while stronger bonds only break after longer and heavier stress conditions.

only the relaxation part needs to be refined. The point to make here is that the 1/6 exponent during the RD stress phase is a result of a delicate interplay between the forward and backward reactions (2). Without the backward reaction, which dominates the time evolution by inserting the 'diffusion-limited' component into the RD model, the forward reaction alone would result in $n = 1$. It is only during relaxation, where the forward rate is suppressed, that the poor performance of the RD reverse reaction becomes visible. Consequently, we do not see any reason to believe the very same reverse reaction to be valid during the stress phase to constructively change the reaction-limited exponent of $n = 1$ to the 'correct' diffusion-limited value of $n = 1/6$.

In a nutshell, we finally have to conclude that reaction-diffusion models cannot capture NBTI. In particular, it appears that hydrogen diffusion does not control the degradation, at least not in the way suggested so far.

Dispersive-Reaction-Rate Models

The models that have been discussed so far assume that there is a single binding energy between the interface state and its passivating hydrogen atom. In fact, there is experimental evidence that this is not the case (54). Rather, due to the amorphous nature of the interfacial layer between the silicon and oxide region the binding energy shows a Gaussian broadening with a variance of about 0.1 eV, see Fig. 17. This dispersion in the binding energy has been used to generalize the forward rate of reaction-diffusion theory (17, 55) by using a suitable average. Since several charge-pumping (CP) experiments indicate that the recovery of interface states is not as pronounced as the recovery of the total ΔV_{th} shift, the generated interface states were assumed to be permanent and the backward reaction was omitted (17, 55).

Although models assuming a dispersion of the defect creation rate could be considered special cases of the RD theory, they are markedly different as in these models diffusion plays no role. These models have been shown to be able to reproduce data obtained by charge-pumping measurements during both stress and recovery (17, 55). In order to explain the significant recovery in the total threshold-voltage shift ΔV_{th}, hole trapping models have been added.

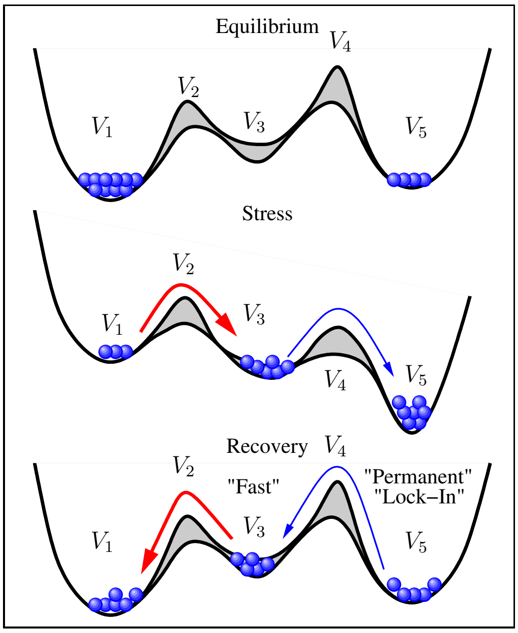

Figure 18: Extension of the dispersive-bond breakage model using three distinct energy levels results in a triple-well model. The second well is energetically higher than the first and the third well and forms a transitional saddle point. Transitions from the second well back to the first well are fast, while the third well represents the permanent component/lock-in.

The Triple-Well Model

The dispersive bond-breaking model was developed based on the observation that charge-pumping currents show poor recovery compared to the overall recovery of ΔV_{th}. This has often been interpreted as being due to the delay: charge-pumping measurements are inherently slow (seconds range) and may therefore miss a considerable fraction of interface-state recovery (18). Secondly, it has been pointed out that CP is highly invasive due to the application of positive bias and may thus distort the measurement results (56).

In order to quickly capture the density of interface states during NBTI degradation without allowing for too much recovery, the on-the-fly charge pumping method has been recently developed (57). Li *et al.* tried to minimize relaxation by using the stress gate voltage as the charge-pumping base level, and pulses with a duty cycle as low as possible. In addition, prior to the actual charge pumping sequence, the dependence of relaxation on the duty cycle of the charge pumping pulses is determined, and subsequent measurements are extrapolated to zero duty cycle.

Based on the results obtained by this method it has been suggested that interface states may also show rapid recovery, a feature conventionally explained by hole detrapping. In an attempt to develop a model that describes NBTI solely by slow and fast recovery of interface states only, we have recently proposed a triple-well model for NBTI (58). The model is based on the assumption that interface state depassivation occurs via an intermediate state into a complete removal of the hydrogen atom, see Fig. 18. Via this intermediate state fast recovery can be explained. Although the model can be fitted to a large amount of data (cf. Fig. 19), there are a number of issues: first, the variance of the intermediate state had to be set to very large values (1 eV) in order to match the data in an extended temperature and voltage range. Such a large variance would be unusual for binding energies and distort the initially assumed Gaussian distribution of binding energies to an effectively flat distribution. Finally, a theoretical and experimental analysis of the

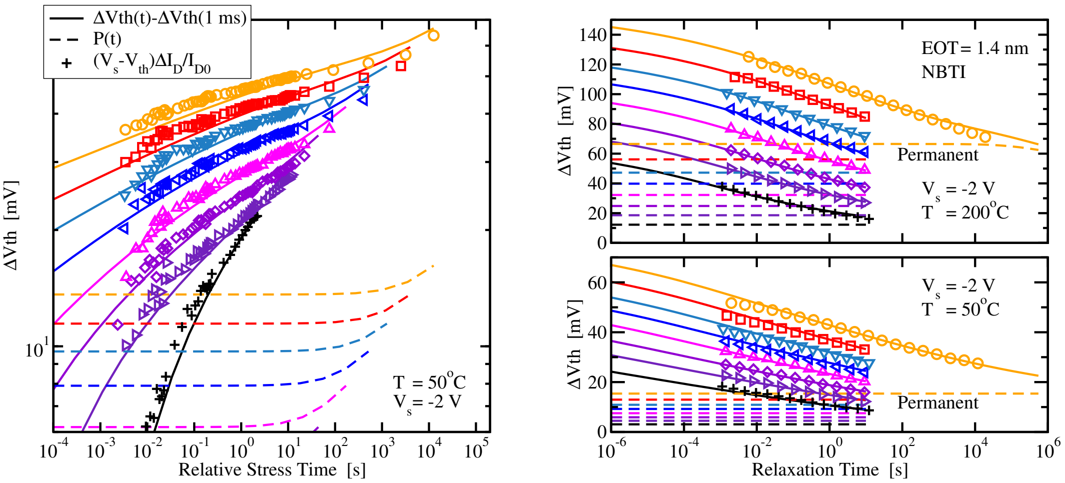

Figure 19: Evaluation of the triple-well model (solid lines: total, dashed lines: ’permanent’ component) against the data taken on SiON devices (58). For the model evaluation the measurement sequence was simulated using fixed parameters in a single simulation run, just like the real measurement. Very good accuracy is obtained for both temperatures during stress (left, shown for 50 °C) and relaxation (right).

on-the-fly charge-pumping technique suggests that the underlying microscopic assumption of fast interface state recovery is very likely an artifact of the method (59).

Consequently, the microscopic interpretation of fast interface state recovery had to be abandoned. Nonetheless, the successful mathematical framework of dispersive rate equations can be used to describe hole trapping in a multiphonon hole capture process (60), which is broadly consistent with existing $1/f$ noise models (61, 62).

Hole Trapping Models

A number of authors have suggested that the pronounced recovery of ΔV_{th} upon the removal of stress is due to hole detrapping, a process consistent with the recovery following a ubiquitous logarithmic time dependence (8). Furthermore, the strong bias-dependence seems to be intuitively compatible with hole trapping and detrapping.

In order to explain the broad distribution of time scales and the bias dependence observed during both stress and recovery, various hole trapping models have been suggested. However, a critical analysis reveals that, although some of these models can produce excellent fits under certain circumstances, see Fig. 20, their underlying microscopic explanation is either missing or questionable. For instance, the detailed hole trapping model developed by Tewksbury (17, 34) is based on elastic hole trapping into pre-existing traps, which in modern ultra-thin gate dielectric layers gives maximum time constants in the millisecond regime only.

Furthermore, elastic hole trapping is only weakly temperature-dependent, in contradiction to short time stress data which can show a pronounced temperature activation and is absent at low temperatures. Finally, elastic hole tunneling would be linearly dependent on the electric field, in contrast to experimental data which show a power-law or exponential dependence on the stress field (60).

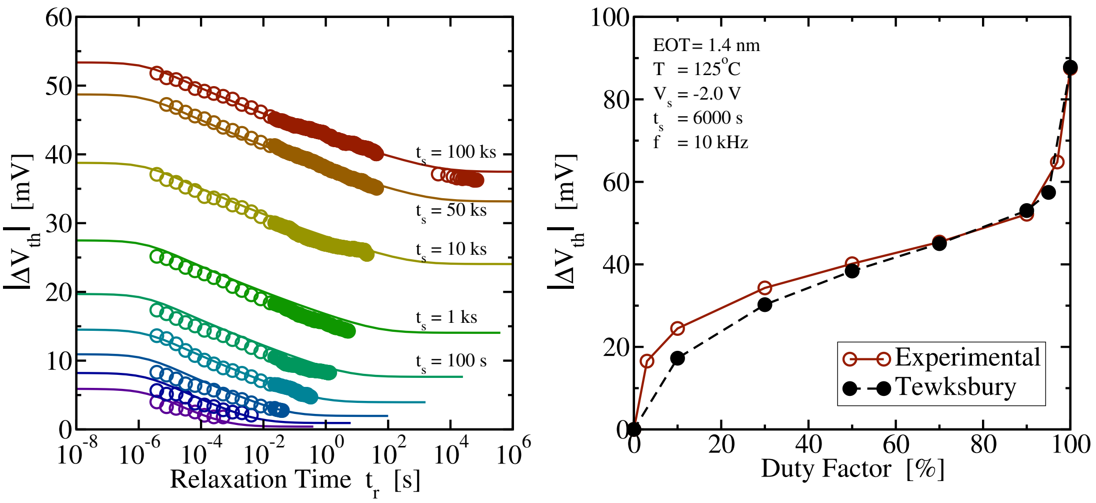

Figure 20: Hole-trapping on top of dispersive defect generation can produce excellent results at a fixed temperature and stress voltage. **Left**: Comparison of recovery data taken after different stress times to experimental data which gives an excellent fit. **Right**: The measured duty factor dependence of (15) can also be excellently reproduced. We remark that such a perfect agreement is only possible under fixed stress conditions and the model parameters would have to be adjusted once e.g. the temperature is changed in order to match our data. Also, the simulation does not account for the 2 ms delay used to take the data of (15), which would smoothen out the simulation result.

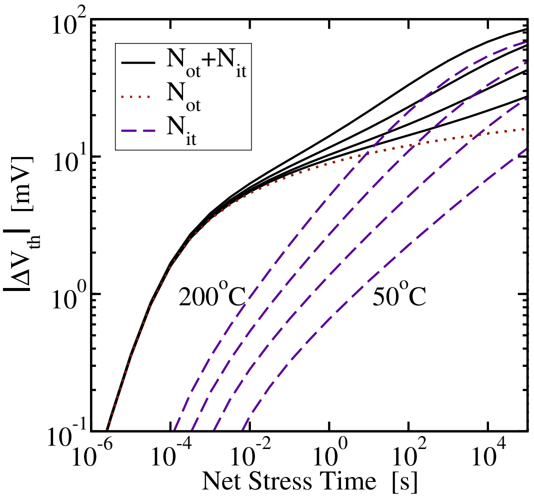

Figure 21: Assuming an independent hole-trapping component on top of interface state generation cannot account for the correct overall temperature dependence of NBTI. In particular, the long-term power-law slope becomes temperature-dependent, in contrast to experimental data.

Combined Models

Reaction-diffusion theory and dispersive-reaction-rate models are frequently combined with hole trapping models to improve the quality of the prediction (17, 67). However, these models do not take the frequently observed correlation between the created interface states and the oxide charges into account (4). In consequence, they often fail to reproduce the temperature and voltage dependence of the overall degradation behavior as the strong temperature dependence of the permanent component results in a 'run-away' with respect to the hole-trapping component, see Fig. 21.

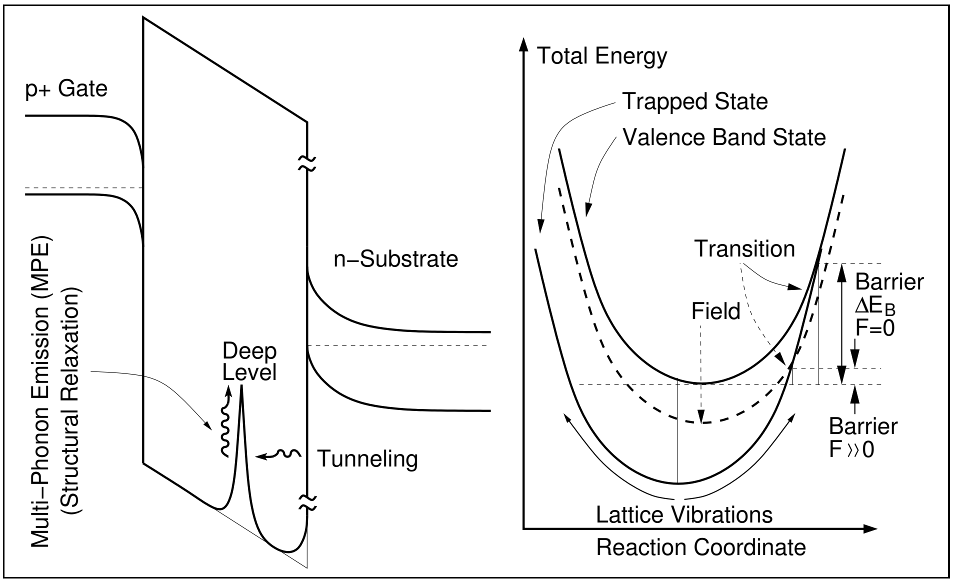

Figure 22: The multiphonon-field-assisted tunneling process used to explain the experimental data: elastic tunneling into deep states is only allowed when the excess energy of holes can be released via a multiphonon emission process. The probability for a thermionic transition over the barrier ΔE_B has been estimated as $\exp(-\Delta E_\mathrm{B}/\mathrm{k_B}T)$ using 1D reaction-coordinate calculations (63–65). Application of an electric field shifts the total energy of the valence band state (dashed line), increasing the transition probability by $\exp(F^2/F_\mathrm{c}^2)$, with F_c being a reference field. (65, 66).

The Two-Stage Model

To overcome the above mentioned issues, we have recently suggested a model where holes are inelastically trapped into deep states which in turn acts as a catalyst to interface state generation (60). The assumptions leading to the model are as follows: under the electric field applied during bias temperature stress, defect precursors in the oxide (oxygen vacancies) break and create hole traps (E' centers). Defect creation and charge trapping can be described using an inelastic multiphonon process (65, 66) as schematically illustrated in Fig. 22. Fundamental arguments from statistical mechanics suggest that the mere presence of these E' centers, which are basically silicon dangling bonds inside the oxide, strongly favors the depassivation of interface states (68). Consequently, oxide and interface charge creation proceeds via two stages and both creation processes are coupled. We remark that such two-stage models provide the standard explanation for defect creation following irradiation (40).

A first evaluation of the model to stress and recovery data, arbitrary bias switches in pure silicon dioxide, nitrided oxides, and well-behaved high-κ gate stacks delivered promising results (60) and calibration results for ultra-thin SiON devices are given in Fig. 23.

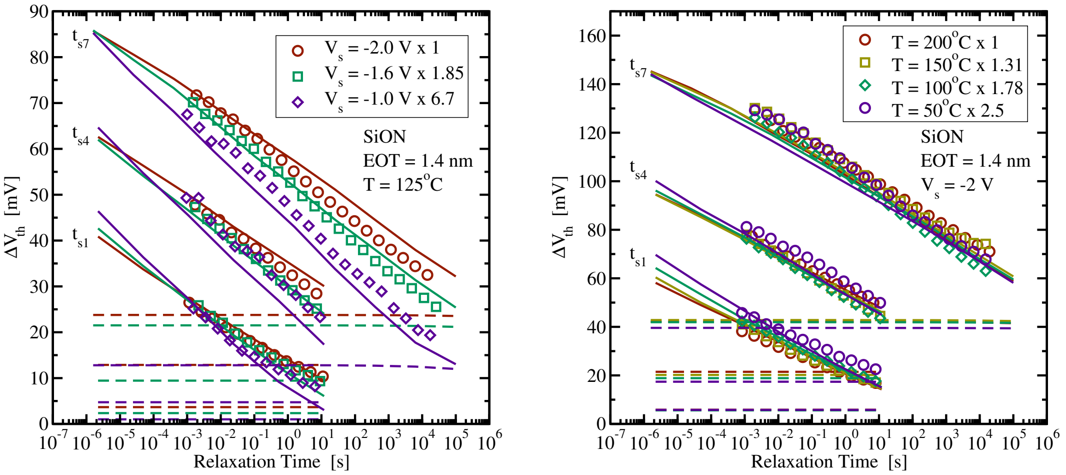

Figure 23: **Left**: Experimental validation of the two-stage model: the degradation resulting from three different bias conditions is compared to the measurement data. The data are given by symbols while the solid lines are the overall simulated degradation with the dashed lines giving the simulated interface state density. For better visibility, the data are normalized to the simulated recovery after the last stress phase at $t_\mathrm{r} = 2\,\mu$s. Clearly, the decrease in slope for increasing stress is well reproduced. **Right**: Temperature scaling: The model also scales properly at different temperatures, avoiding the 'run-away' of the permanent component present in the uncoupled model.

Conclusions

We have thoroughly analyzed existing NBTI models and identified a number of serious shortcomings, implying that the physical assumptions underlying these models cannot be correct. Most notably, all models using some hydrogen-diffusion control must be ruled out in their present forms. The conventionally used hole trapping models, on the other hand, are temperature-independent and show a linear field dependence, in contrast to data. Based on these results we suggest a new model using multiphonon-field-assisted tunneling which delivers promising results.

References

1. K. Jeppson, and C. Svensson, *J.Appl.Phys.* **48**, 2004–2014 (1977).
2. M. Alam, H. Kufluoglu, D. Varghese, and S. Mahapatra, *Microelectronics Reliability* **47**, 853–862 (2007).
3. C. Shen, M.-F. Li, X. Wang, Y.-C. Yeo, and D.-L. Kwong, *IEEE Electron Device Lett.* **27**, 55–57 (2006).
4. V. Huard, M. Denais, and C. Parthasarathy, *Microelectronics Reliability* **46**, 1–23 (2006).
5. 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_th-Measurements," in *Proc. Intl.Rel.Phys.Symp.*, 2006, pp. 448–453.
6. T. Grasser, W. Goes, V. Sverdlov, and B. Kaczer, "The Universality of NBTI Relaxation and its Implications for Modeling and Characterization," in *Proc. Intl.Rel.Phys.Symp.*, 2007, pp. 268–280.

7. A. Haggag, G. Anderson, S. Parihar, D. Burnett, G. Abeln, J. Higman, and M. Moosa, "Understanding SRAM High-Temperature-Operating-Life NBTI: Statistics and Permanent vs Recoverable Damage," in *Proc. Intl.Rel.Phys.Symp.*, 2007, pp. 452–456.
8. B. Kaczer, T. Grasser, P. Roussel, J. Martin-Martinez, R. O'Connor, B. O'Sullivan, and G. Groeseneken, "Ubiquitous Relaxation in BTI Stressing-New Evaluation and Insights," in *Proc. Intl.Rel.Phys.Symp.*, 2008, pp. 20–27.
9. D. Ang, S. Wang, G. Du, and Y. Hu, *IEEE Trans.Dev.Mat.Rel.* **8**, 22–34 (2008).
10. 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. Intl.Rel.Phys.Symp.*, 2005, pp. 381–387.
11. M. Denais, A. Bravaix, V. Huard, C. Parthasarathy, G. Ribes, F. Perrier, Y. Rey-Tauriac, and N. Revil, "On-the-fly Characterization of NBTI in Ultra-Thin Gate Oxide pMOSFET's," in *Proc. Intl.Electron Devices Meeting*, 2004, pp. 109–112.
12. A. Islam, E. N. Kumar, H. Das, S. Purawat, V. Maheta, H. Aono, E. Murakami, S. Mahapatra, and M. Alam, "Theory and Practice of On-the-fly and Ultra-fast V_T Measurements for NBTI Degradation: Challenges and Opportunities," in *Proc. Intl.Electron Devices Meeting*, 2007, pp. 1–4.
13. H. Reisinger, U. Brunner, W. Heinrigs, W. Gustin, and C. Schlünder, *IEEE Trans.Dev.Mat.Rel.* **7**, 531–539 (2007).
14. T. Grasser, P.-J. Wagner, P. Hehenberger, W. Goes, and B. Kaczer, *IEEE Trans.Dev.Mat.Rel.* **8**, 526 – 535 (2008).
15. T. Grasser, B. Kaczer, P. Hehenberger, W. Goes, R. O'Connor, H. Reisinger, W. Gustin, and C. Schlünder, "Simultaneous Extraction of Recoverable and Permanent Components Contributing to Bias-Temperature Instability," in *Proc. Intl.Electron Devices Meeting*, 2007, pp. 801–804.
16. R. Fernandez, B. Kaczer, A. Nackaerts, S. Demuynck, R. Rodriguez, M. Nafria, and G. Groeseneken, "AC NBTI Studied in the 1 Hz - 2 GHz Range on Dedicated On-Chip CMOS Circuits," in *Proc. Intl.Electron Devices Meeting*, 2006, pp. 1–4.
17. V. Huard, C. Parthasarathy, N. Rallet, C. Guerin, M. Mammase, D. Barge, and C. Ouvrard, "New Characterization and Modeling Approach for NBTI Degradation from Transistor to Product Level," in *Proc. Intl.Electron Devices Meeting*, 2007, pp. 797–800.
18. S. Mahapatra, K. Ahmed, D. Varghese, A. E. Islam, G. Gupta, L. Madhav, D. Saha, and M. A. Alam, "On the Physical Mechanism of NBTI in Silicon Oxynitride p-MOSFETs: Can Differences in Insulator Processing Conditions Resolve the Interface Trap Generation versus Hole Trapping Controversy?," in *Proc. Intl.Rel.Phys.Symp.*, 2007, pp. 1–9.
19. H. Reisinger, O. Blank, W. Heinrigs, W. Gustin, and C. Schlünder, *IEEE Trans.Dev.Mat.Rel.* **7**, 119–129 (2007).
20. J. Zhang, Z. Ji, M. Chang, B. Kaczer, and G. Groeseneken, "Real V_th Instability of pMOSFETs Under Practical Operation Conditions," in *Proc. Intl.Electron Devices Meeting*, 2007, pp. 817–820.
21. IμE, *MINIMOS-NT 2.0 User's Guide*, Institut für Mikroelektronik, Technische Universität Wien, Austria (2002), http://www.iue.tuwien.ac.at/software/minimos-nt.
22. M. Denais, A. Bravaix, V. Huard, C. Parthasarathy, C. Guerin, G. Ribes, F. Perrier, M. Mairy, and D. Roy, "Paradigm Shift for NBTI Characterization in Ultra-Scaled CMOS Technologies," in *Proc. Intl.Rel.Phys.Symp.*, 2006, pp. 735–736.

23. T. Grasser, and B. Kaczer, “Negative Bias Temperature Instability: Recoverable versus Permanent Degradation,” in *Proc. ESSDERC*, 2007, pp. 127–130.
24. T. Grasser, W. Goes, and B. Kaczer, “Towards Engineering Modeling of Negative Bias Temperature Instability,” in *Defects in Microelectronic Materials and Devices*, edited by D. Fleetwood, R. Schrimpf, and S. Pantelides, Taylor and Francis/CRC Press, 2008, pp. 1–30.
25. T. Grasser, “Negative Bias Temperature Instability: Modeling Challenges and Perspectives,” in *Proc. Intl.Rel.Phys.Symp.*, 2008, (Tutorial).
26. W. Jackson, and C. Tsai, *Physical Review B* **45**, 6564–6580 (1992).
27. N. Nickel, W. Jackson, and J. Walker, *Physical Review B* **53**, 7750–7761 (1996).
28. M. Alam, “Negative Bias Temperature Instability: Basics/Modeling,” in *Proc. Intl.Rel.Phys.Symp.*, 2005, (Tutorial).
29. M. Alam, “A Critical Examination of the Mechanics of Dynamic NBTI for pMOSFETs,” in *Proc. Intl.Electron Devices Meeting*, 2003, pp. 345–348.
30. S. Chakravarthi, A. Krishnan, V. Reddy, C. Machala, and S. Krishnan, “A Comprehensive Framework for Predictive Modeling of Negative Bias Temperature Instability,” in *Proc. Intl.Rel.Phys.Symp.*, 2004, pp. 273–282.
31. M. Alam, and H. Kufluoglu, *ECS Trans.* **1**, 139–145 (2005).
32. A. Krishnan, C. Chancellor, S. Chakravarthi, P. Nicollian, V. Reddy, A. Varghese, R. Khamankar, and S. Krishnan, “Material Dependence of Hydrogen Diffusion: Implications for NBTI Degradation,” in *Proc. Intl.Electron Devices Meeting*, 2005, pp. 688–691.
33. H. Kufluoglu, and M. Alam, *IEEE Trans.Electron Devices* **54**, 1101–1107 (2007).
34. T. Tewksbury, *Relaxation Effects in MOS Devices due to Tunnel Exchange with Near-Interface Oxide Traps*, Ph.D. Thesis, MIT (1992).
35. J. Kakalios, R. A. Street, and W. B. Jackson, *Physical Review Letters* **59**, 1037–1040 (1987).
36. E. Montroll, and H. Scher, *J.Stat.Phys* **9**, 101–135 (1973).
37. H. Scher, and E. Montroll, *Physical Review B* **12**, 2455–2477 (1975).
38. J. Noolandi, *Physical Review B* **16**, 4466–4473 (1977).
39. V. Arkhipov, and A. Rudenko, *Philos.Mag.B* **45**, 189–207 (1982).
40. D. Brown, and N. Saks, *J.Appl.Phys.* **70**, 3734–3747 (1991).
41. J. Orenstein, M. Kastner, and V. Vaninov, *Philos.Mag.B* **46**, 23–62 (1982).
42. W. Jackson, *Physical Review B* **38**, 3595–3598 (1988).
43. V. Arkhipov, “Trap-Controlled and Hopping Modes of Transport and Recombination: Similarities and Differences,” in *Proc. Intl.Symp.Elect.Insul.Materials*, 1995, pp. 271–274.
44. J. Noolandi, *Physical Review B* **16**, 4474–4479 (1977).
45. F. Schmidlin, *Physical Review B* **16**, 2362–2385 (1977).
46. M. Alam, and S. Mahapatra, *Microelectronics Reliability* **45**, 71–81 (2005).
47. M. Houssa, M. Aoulaiche, S. D. Gendt, G. Groeseneken, M. Heyns, and A. Stesmans, *Appl.Phys.Lett.* **86**, 1–3 (2005).
48. S. Zafar, *J.Appl.Phys.* **97**, 1–9 (2005).
49. B. Kaczer, V. Arkhipov, R. Degraeve, N. Collaert, G. Groeseneken, and M. Goodwin, *Appl.Phys.Lett.* **86**, 1–3 (2005).
50. T. Grasser, W. Goes, and B. Kaczer, *IEEE Trans.Dev.Mat.Rel.* **8**, 79–97 (2008).
51. F. McLean, and G. Ausman, *Physical Review B* **15**, 1052–1061 (1977).

52. T. Grasser, R. Entner, O. Triebl, H. Enichlmair, and R. Minixhofer, "TCAD Modeling of Negative Bias Temperature Instability," in *Proc. Simulation of Semiconductor Processes and Devices*, Monterey, USA, 2006, pp. 330–333.
53. D. Varghese, D. Saha, S. Mahapatra, K. Ahmed, F. Nouri, and M. Alam, "On the Dispersive versus Arrhenius Temperature Activation of NBTI Time Evolution in Plasma Nitrided Gate Oxides: Measurements, Theory, and Implications," in *Proc. Intl.Electron Devices Meeting*, 2005, pp. 1–4.
54. A. Stesmans, *Physical Review B* **61**, 8393–8403 (2000).
55. A. Haggag, W. McMahon, K. Hess, K. Cheng, J. Lee, and J. Lyding, "High-Performance Chip Reliability from Short-Time-Tests," in *Proc. Intl.Rel.Phys.Symp.*, 2001, pp. 271–279.
56. B. Kaczer, T. Grasser, R. Fernandez, and G. Groeseneken, "Toward Understanding the Wide Distribution of Time Scales in Negative Bias Temperature Instability," in *Silicon Nitride, Silicon Dioxide, and Emerging Dielectrics 9*, edited by R. Sah, J. Zhang, Y. Kamakura, M. Deen, and J. Yota, ECS Transactions, 2007, vol. 6, pp. 265–281.
57. M.-F. Li, D. Huang, C. Shen, T. Yang, W.J., W. Liu, and Z. Liu, *IEEE Trans.Dev.Mat.Rel.* **8**, 62–71 (2008).
58. T. Grasser, B. Kaczer, and W. Goes, "An Energy-Level Perspective of Bias Temperature Instability," in *Proc. Intl.Rel.Phys.Symp.*, 2008, pp. 28–38.
59. P. Hehenberger, T. Aichinger, T. Grasser, W. Goes, O. Triebl, B. Kaczer, and M. Nelhiebel, "Do NBTI-Induced Interface States Show Fast Recovery? A Study Using a Corrected On-The-Fly Charge-Pumping Measurement Technique," in *Proc. Intl.Rel.Phys.Symp.*, 2009, (in print).
60. T. Grasser, B. Kaczer, W. Goes, T. Aichinger, P. Hehenberger, and M. Nelhiebel, "A Two-Stage Model for Negative Bias Temperature Instability," in *Proc. Intl.Rel.Phys.Symp.*, 2009, (in print).
61. A. McWhorter, *Sem.Surf.Phys* pp. 207–228 (1957).
62. M. Kirton, and M. Uren, *Adv.Phys* **38**, 367–486 (1989).
63. D. Lang, and C. Henry, *Physical Review Letters* **35**, 1525–1528 (1975).
64. C. Henry, and D. Lang, *Physical Review B* **15**, 989–1016 (1977).
65. S. Ganichev, W. Prettl, and I. Yassievich, *Phys.Solid State* **39**, 1703–1726 (1997).
66. S. Makram-Ebeid, and M. Lannoo, *Physical Review B* **25**, 6406–6424 (1982).
67. A. Islam, H. Kufluoglu, D. Varghese, S. Mahapatra, and M. Alam, *IEEE Trans.Electron Devices* **54**, 2143–2154 (2007).
68. P. Lenahan, *Microelectronic Engineering* **69**, 173–181 (2003).