THE UNIVERSALITY OF NBTI RELAXATION AND ITS IMPLICATIONS FOR MODELING AND CHARACTERIZATION

Tibor Grasser*, Wolfgang Gös*, Victor Sverdlov†, and Ben Kaczer°

* Christian Doppler Laboratory for TCAD at the †Institute for Microelectronics, TU Wien, Gusshausstr. 27–29, A-1040 Wien, Austria

° IMEC, Kapeldreef 75, B-3001 Leuven, Belgium

ABSTRACT

As of date many NBTI models have been published which aim to successfully capture the essential physics [1–5]. As such, these models have mostly focused on the stress phase. The relaxation phase, on the other hand, has not received as much attention, possibly because of the contradictory results published so far. Particularly noteworthy are the very long relaxation tails of almost logarithmic nature [5–7], which cannot be successfully described by the reaction-diffusion model [7]. We argue that understanding the nature of the relaxation phase could hold the key to unraveling the underlying NBTI mechanism. In particular, we stipulate that the relaxation phase follows a universal relaxation 'law' [6, 8–10], demonstrate the valuable consequences resulting therefrom, and use this universality to classify presently available NBTI models.

I. INTRODUCTION

Although being known for forty years [11], negative bias temperature instability (NBTI) is attracting an ever growing industrial and scientific attention as one of the most important reliability issues in modern CMOS technology. NBTI is mostly described by a shift of the threshold voltage when a typically large negative voltage is applied to the gate of a pMOS at elevated temperatures [12, 13]. In the following we assume that during NBT stress the change in the density of interface states, which are commonly assumed to be P_b centers [14], is given through $\Delta N_\mathrm{it}(t)$. It is assumed that charging and discharging of these interface states is very fast, and consequently that the positive charge in these interface states follows the Fermi-level via

$$\Delta Q_\mathrm{it}(t) = \mathrm{q} \int \Delta D_\mathrm{it}(E_\mathrm{t},t)\, f(E_\mathrm{F},E_\mathrm{t},t)\, \mathrm{d}E_\mathrm{t} \ . \tag{1}$$

Here, ΔD_it is the time-dependent density of interface states, which is by a still to be quantified relation directly linked to $\Delta N_\mathrm{it}(t)$ [15], and $f(E_\mathrm{t})$ their occupancy. During NBT stress, the Fermi-level E_F is close to the valence band edge and $f(E_\mathrm{t}) \approx 1$ throughout the silicon bandgap. Thus, during stress, under the assumption that P_b-centers introduce states only *within* the silicon bandgap, see [4] for a different interpretation, all newly generated interface states ΔN_it are positively charged and one obtains independently of the exact form of the density-of-states $\Delta Q_\mathrm{it}(t) \approx \mathrm{q}\Delta N_\mathrm{it}(t)$. This is the usual assumption employed for instance in the reaction-diffusion model.

On top of generated interface defects, charge may be stored in existing or newly created oxide traps. Although most of these traps may still be considered 'fast', they are more difficult to charge and discharge, that is, have larger time constants than interface states due to their location inside the oxide bulk. Thus, their occupancy cannot follow the Fermi-level and $\Delta Q_\mathrm{ox}(t)$ will be governed by different dynamics. The contribution of the oxide charges to the threshold voltage shift is formally written as

$$\Delta Q_\mathrm{ox}(t) = \mathrm{q} \iint \Delta D_\mathrm{ox}(x,E_\mathrm{t},t)\, f_\mathrm{ox}(x,E_\mathrm{t},t)\, (1 - x/t_\mathrm{ox})\, \mathrm{d}x\, \mathrm{d}E_\mathrm{t} \ , \tag{2}$$

with ΔD_ox being the spatially and energy-dependent density-of-states in the oxide, f_ox the occupancy of these traps, and t_ox the oxide thickness. Note that the issue of whether oxide charges are important during NBTI or not is one of the most controversial at the time [2, 4, 16, 17]. Also, the question whether ΔD_ox consists mainly of pre-existing traps [2, 16] or traps that are created during stress [4] remains to be answered.

Other potential contributions to a threshold voltage shift like mobile charges are commonly assumed to be negligible in the context of NBTI and the total threshold voltage shift is thus given by

$$\Delta V_\mathrm{th}(t) = \frac{\Delta Q_\mathrm{it}(t) + \Delta Q_\mathrm{ox}(t)}{C_\mathrm{ox}} \ . \tag{3}$$

During stress, most measurements indicate that $\Delta V_\mathrm{th}(t)$ follows a power-law as At^n [12]. However, log-like behavior, in particular at short times, has also been reported [7, 18, 19].

The fundamental problem in the context of NBTI is given by the fact that the damage created during the stress phase begins to recover immediately once the stress is removed. This makes the classic measurement technique where the stress is interrupted during the extraction of the threshold voltage problematic [7, 20]. In particular, the value of the extracted power-law exponent depends significantly on the delay introduced during the measurement [5, 17, 21]. Experimental results obtained with delayed measurements show a linear increase of the exponent with temperature [5, 17, 22] with values around $0.2 - 0.3$. In contrast, temperature-independent exponents in the range $0.07 - 0.2$ have been extracted from recent delay-free measurements [9, 19, 23]. Of particular interest is the question related to the origin of this relaxation. While some authors assume that hole trapping is negligible and both degradation as well as relaxation is determined by the temporal change of the interface state density [17], others acknowledge at least partial importance of trapped charges [4, 19, 22, 24]. In the latter case it has been assumed that trapped charges either form the 'fast' component of NBTI relaxation superimposed onto some interface defect relaxation [19, 24] or are solely responsible for any recovery while created interface defects do not recover at all [4, 22].

II. MEASUREMENT ISSUES

The understanding and characterization of NBTI is considerably hampered by the difficulties arising during measurement. Currently, two techniques are used to characterize NBTI: the classic measurement/stress/measurement (MSM) technique, which is handicapped by undesired relaxation, and on-the-fly (OTF) measurements which avoid any relaxation by maintaining a high stress level throughout the measurement and directly monitor the drain current in the linear regime, $\Delta I_\mathrm{D,lin}$. Since it is the threshold voltage shift ΔV_th rather than $\Delta I_\mathrm{D,lin}$ that is relevant for design purposes, $\Delta I_\mathrm{D,lin}$ has to be converted into the more relevant ΔV_th which involves some approximate relations [22] or an empirical formalism [25]. This issue is of particular importance during the assessment of the relaxation phase: When V_G is

1-4244-0919-5/07/$25.00 ©2007 IEEE

IEEE 07CH37867 45th Annual International Reliability Physics Symposium, Phoenix, 2007

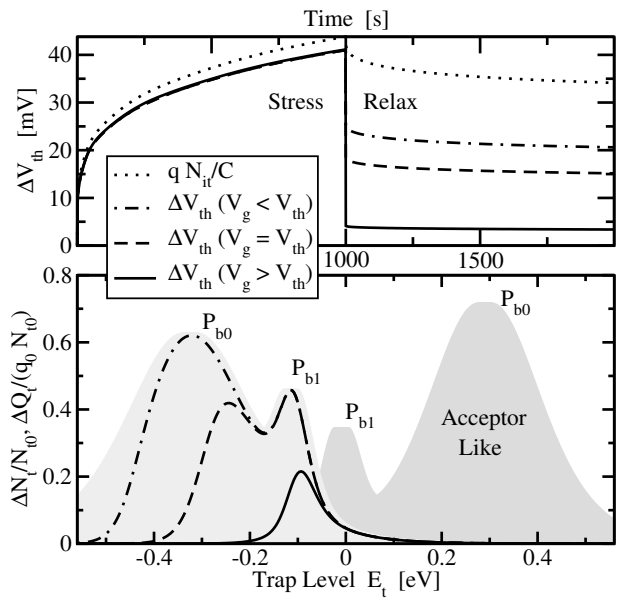

Fig. 1: Influence of the interface state occupancy on the observed threshold voltage shift during OTF measurements. During stress, nearly all interface traps are positively charged. When a different gate voltage is used during relaxation, only a fraction of the traps are visible which must be separated from the real relaxation. Schematically shown is the density-of-states typically associated with P_{b0} and P_{b1} centers [26].

left at V_G^{relax} [6], the interface trap occupancy is considerably lower than during the stress phase [15], resulting in spurious additional relaxation (Fig. 1). Conversely, when V_G is brought back to V_G^{stress} [9], one faces the opposite problem one is trying to avoid during the stress phase, since now additional uncontrolled stress is introduced during the measurement cycles. Even more important is the fact that the initial value of $I_{D,lin}$ is extremely difficult to determine as it is already obtained at the stress voltage. Conventionally, the time required for this is in the milliseconds range where already significant degradation can be observed [19] but any uncertainty in $I_{D,lin}$ modifies the time exponent (the 'slope') of ΔV_{th} on a log-log plot. This may render many results obtained by the OTF technique questionable. In contrast, the MSM technique probes the interface under comparable conditions during both the stress and relaxation phase. In addition, the voltage applied to the gate is close to the threshold-voltage where only negligible degradation can be expected. However, as is shown below, it is probably **very difficult to minimize the measurement delay** in such a way that the true degradation is observed.

III. Characterization of Relaxation

To formally distinguish between the degradation during the stress and relaxation phases we use the term $S(t_s)$ for the real damage accumulated during the stress phase. As soon as the stress voltage is removed, relaxation sets in as a function of the accumulated stress time t_s and the relaxation time $t_r = t - t_s$, which will be denoted as $R(t_s, t_r)$. Furthermore, we introduce $S_M(t_s, t_M)$ as the observed damage during an MSM sequence with a measurement delay of t_M.

Since the damage $S(t_s)$ is known to relax as soon as the stress voltage is removed, possible at timescales shorter than a microsecond [7], a rigorous characterization of the relaxation phase is extremely challenging. Typically, the relaxation data $R(t_s, t_r)$ recorded at different stress times t_s have been normalized to the first measurement point t_M as $r_f(t_s, t_r) = R(t_s, t_r)/R(t_s, t_M)$, the *fractional recovery*, and aligned

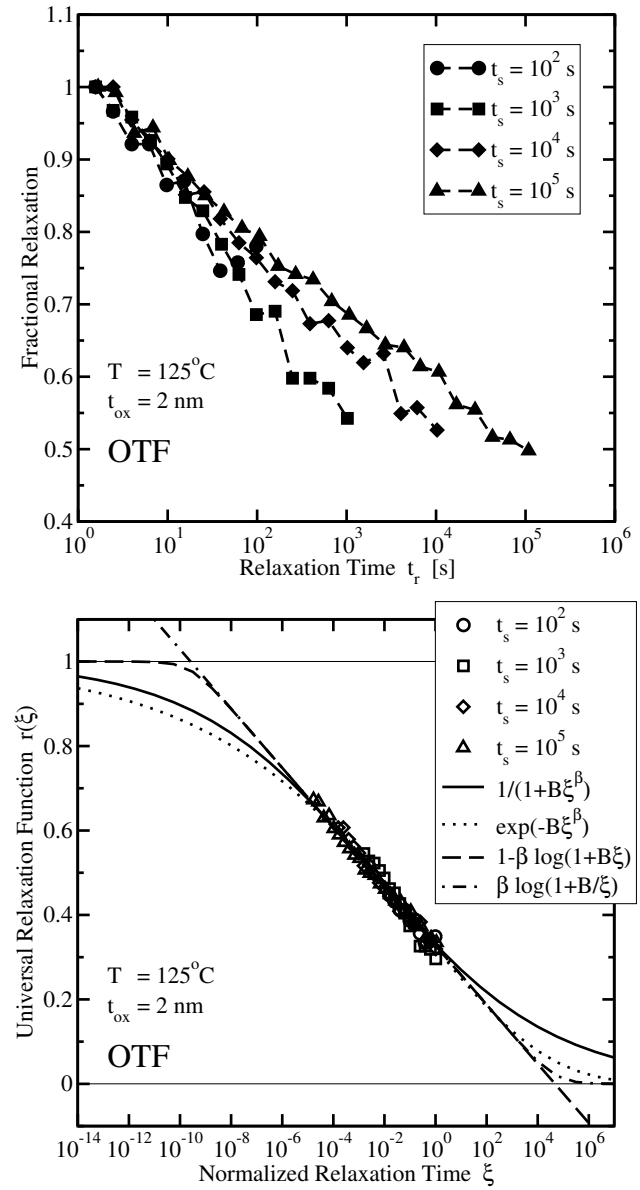

Fig. 2: Demonstration of universal recovery for the OTF data of Denais *et al.* [10]. The top figure shows a conventional view of the fractional recovery as a function of the relaxation time t_r. Apparently, data obtained after longer stress times, seems to relax more slowly than data obtained at shorter times. The bottom figure, on the other hand, demonstrates the universality of relaxation when the relaxation data is normalized to the last stress value and plotted over the ratio $\xi = t_r/t_s$ [10]. Also shown are some possible empirical expressions which can be fit to the data.

as a function of the absolute relaxation time t_r [5–8], see Fig. 2. Although the functional form of the relaxation remains illusive in such a plot, Rangan *et al.* [8] were the first to observe certain features of a universality in the relaxation. This and subsequent studies can be summarized as follows:

R1) Recovery does not level off even at the shortest times [8] which implies that it is difficult to asses the real damage a $t_r = 0$. This is even true for the ultra-fast data obtained by Reisinger *et al.* [7] with $t_M = 2.2\,\mu$s.

R2) For longer stress times, the relaxation apparently slows down [5,

8, 9], that is, the fractional recovery becomes smaller, as visible in Fig. 2.

R3) Independent of the stress field, the fractional recovery is roughly similar [8].

R4) The universality as proposed by Rangan *et al.* has been studied by Krishnan *et al.* [9], who observed that the fractional recovery is even independent of the stress-time. This, as also pointed out by Krishnan *et al.*, is somewhat in disagreement with R2 which states that the fractional recovery becomes smaller with larger stress times.

R5) Recovery seems to reset a certain fraction of the defects to their original state, as proven by re-stress experiments [8, 22].

R6) Denais *et al.* [10] realized that the relative recovery obtained after different stress times follows the same pattern when plotted as a function of relaxation time over stress time, $\xi = t_r/t_s$, see Fig. 2.

R7) Contradicting evidence is available regarding the recovery field-dependence, where [8, 17] report no field-dependence contrary to [5, 22] who observe a strong field-dependence.

As will be shown, the universality described by Denais *et al.* is both the most intriguing and consequential feature as it has a fundamental impact on any theoretical understanding of NBTI and should therefore be an important ingredient in any modeling attempt. Considering also the fact that not all damage can recover, that is, that there could be a permanent component $P(t_s)$ [8, 27], we rewrite the accumulated damage as $S(t_s) = R(t_s, 0) + P(t_s)$ and introduce the universal relaxation function as

$$r(\xi) = \frac{R(t_s, t_r)}{S(t_s) - P(t_s)} = \frac{R(t_s, t_r)}{R(t_s, 0)} \tag{4}$$

and discuss its properties and consequences in the following. Note the relation between the universal recovery function and the fractional recovery given by $r_f(t_s, t_r) = r(\xi)/r(\xi_M)$ with $\xi_M = t_M/t_s$. Since the data available to us does not definitely support the existence of a permanent component, we will assume the permanent component to be negligible in the following and assume $P(t_s) = 0$.

A. Functional Form of the Relaxation

Lacking a universally accepted and valid theory for NBTI, the exact form of the universal relaxation function $r(\xi)$ remains illusive at this point and we will have to empirically estimate $r(\xi)$ in the following. Various empirical expressions have already been suggested and fit to measured relaxation data. Generalized for universality, these expressions include $r(\xi) = 1 - B\xi^\beta$ [28], $r(\xi) = 1 - B\log(\xi)$ [29], and $r(\xi) = 1 - C(1 - \exp(-B\xi))$ [30]. These expressions are problematic as they either have no asymptotic limit for large and/or small ξ, or they do not capture recent measurement data. In particular, we will show that the limit $r(\xi \to 0)$ is of considerable practical value and should be meaningful in any analytic expression.

Interestingly, NBTI models operating in the diffusion-limited regime, be it classic [29] or dispersive diffusion [5], agree that the relaxation depends on the ratio $\xi = t_r/t_s$ only, that is, they are universal in our sense. For the reaction-diffusion (RD) model Alam [29] derived the approximate solution $r(\xi) = 1 - \sqrt{\gamma\xi/(1+\xi)}$ with the empirical parameter $\gamma \approx 0.5$. Although this expression is accurate for small ξ, for large ξ it has the limit $1 - \sqrt{\gamma}$ which for $\gamma \neq 1$ does not agree with the numerical solution of the RD model which exactly goes to zero. It can be shown that the power-law-like expression

$$r(\xi) = 1/(1 + \xi^{1/2}) \tag{5}$$

agrees very well with the numerical solution over the whole range of relaxation times (cf. Fig. 12). Generalizing the RD model to allow for dispersive transport of the hydrogen-species, Kaczer *et al.* [5] derived

$$r(\xi) = 1/(1 + \xi^{\alpha/2}) \tag{6}$$

with α being the dispersion parameter ($\alpha \in [0 \ldots 1]$). Note that in the diffusive limit where $\alpha = 1$, the RD result is retained. However, as will become clear in the sequel, neither (5) nor (6) can cover the whole range of available measurement data and we will use the generalized form

$$r(\xi) = 1/(1 + B\xi^\beta) \tag{7}$$

where the parameters B and β are in the range $B \approx 0.3 - 3$ and $\beta \approx 0.15 - 0.2$ for most of the data available to us.

Since the exact form of the relaxation function has important implications on the interpretation of the measurement data, we also consider alternative expressions. First, we note that relaxation in disordered systems has long been described using a stretched-exponential [31]

$$r(\xi) = \exp(-B\xi^\beta) \ . \tag{8}$$

Also, Denais *et al.* [10] suggested the empirical expression

$$r(\xi) = 1 - \beta\log(1 + B\xi) \ , \tag{9}$$

while Huard *et al.* [18] used the analytic solution of a hole-trapping problem derived in [32] which appears to be of a similar form (see Section VII-C for a discussion)

$$r(\xi) = \beta\log(1 + B/\xi) \ . \tag{10}$$

We note that (9) has no useful limit for large ξ while (10) goes to infinity for $\xi \to 0$. Since we will make extensive use of the latter limit, we will use the empirical relation (9) in our comparisons, noting that the singularity in the physics-based expression (10) is a consequence of the simplifications employed in the derivation presented in [32] and does of course not exist in the full numerical solution. Nevertheless, both expressions are equivalent for intermediate values of ξ and will be labeled 'log-like' in the following.

The expressions summarized above are compared in Fig. 2. All expressions can be fit to the measurement data and give fits of practically the same accuracy. However, they result in different extrapolations for large and small relaxation times, the consequences of which need to be carefully studied.

IV. Characterization of MSM Data

Although more delicate to apply, universal relaxation is of particular interest for data obtained by the MSM technique. For the normalization needed in (4) one has to keep in mind that the value of $S(t_s) = R(t_s, 0)$ is essentially unknown, one only knows $R(t_s, t_M)$ determined at the first measurement point available after a short relaxation period t_M. However, making use of the universal relaxation expression (4) and *assuming* for the time being that $r(\xi)$ is known, $S(t_s) = R(t_s, 0)$ can be obtained as

$$S(t_s) = \frac{R(t_s, t_M)}{r(t_M/t_s)} \ . \tag{11}$$

Inserting the above into the universal relaxation relation (4) we obtain

$$\frac{r(\xi)}{r(\xi_M)} = \frac{R(t_s, t_r)}{R(t_s, t_M)} \ . \tag{12}$$

From (12) the as of yet unknown parameters B and β can be easily determined from a measured sequence of relaxation points $R(t_s, t_r)$,

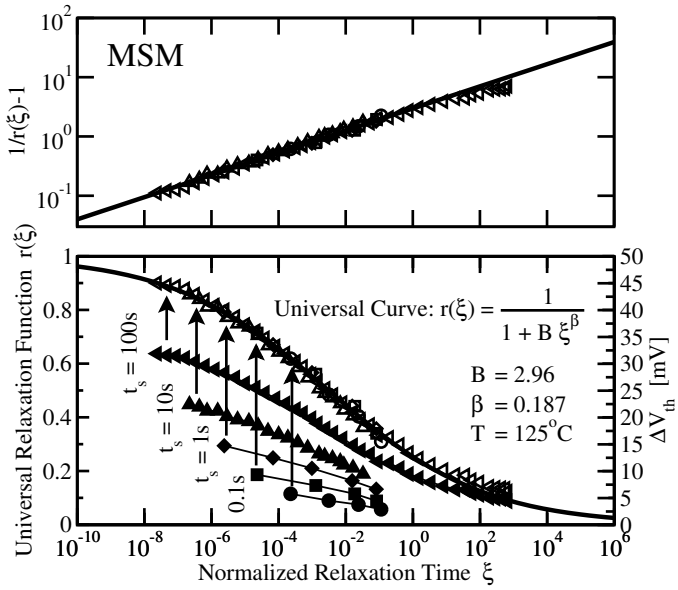

Fig. 3: Application of universal relaxation to the ultra-fast MSM data obtained by Reisinger *et al.* [7]. Depending on the choice of the universal relaxation function, the individual data points can be mapped onto the respective universal curve, in this case (7). Note the linear behavior of $1/r-1$ shown in the upper plot. The slight deviation for $\xi > 10^2$ could bear the hint of a permanent component $P(t_s)$.

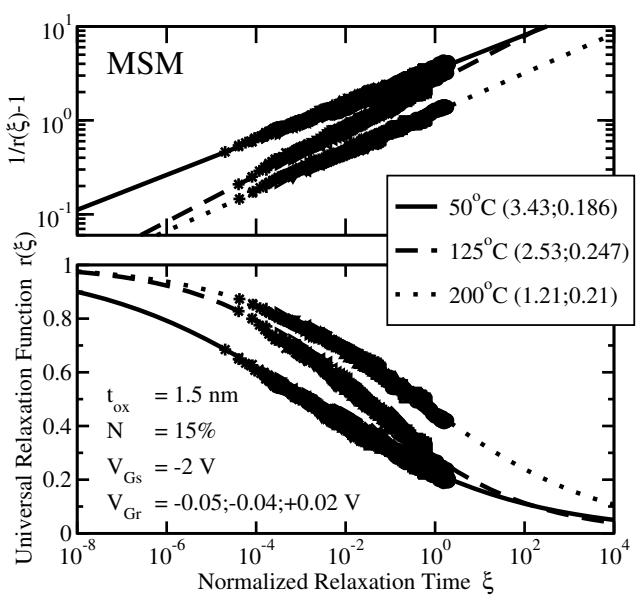

Fig. 4: Same as Fig. 3 but with data from IMEC [5]. Relaxation data of three devices stressed in a single MSM sequence were recorded at 10 different stress-times in the interval $10\,\mathrm{s} - 10^4\,\mathrm{s}$ at three different temperatures. The values of B and β (given in parenthesis) depend on the temperature, β even in a non-monotonic manner which might indicate the existence of two different processes with different temperature dependencies [7].

see for example Fig. 3. Naturally, in contrast to data obtained by OTF measurements where $R(t_s,0)$ is known, the analytical expression determines the final value of $R(t_s,0)$ through the extrapolation given by (12). This results in a 'floating' behavior of $r(\xi_M)$ which reflects the uncertainty of this approach.

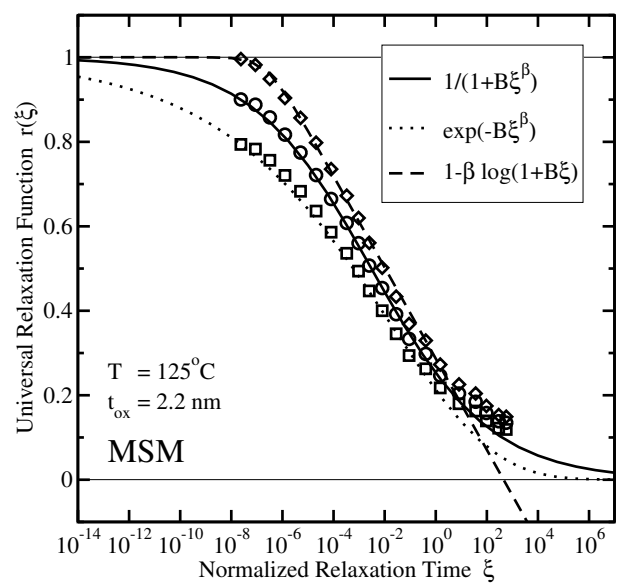

Fig. 5: The uncertainty introduced by extrapolating back to $t_M = 0$ using a particular relaxation function. This uncertainty is in the order of 10% for the relaxation functions considered here. For the sake of clarity, only the $t_s = 100\,\mathrm{s}$ data of Reisinger *et al.* [7] is shown. Note that the different extrapolations to $t_M = 0$ shift the same measurement data (various symbols) by a multiplicative factor.

Before looking at this issue, however, we consider the case of a typical MSM sequence. We recall that during MSM sequences the duration of the stress intervals usually grows exponentially while the measurement interval t_M is short and of constant duration. This implies that after a certain stress time, which we determined empirically to be of the order $t_s \gtrsim 10 \times t_M$, the relaxation during the measurement does not significantly alter the damage at the *end* of each stress phase, meaning that the damage relaxed during each measurement interval is mostly restored during the next stress phase. Consequently, equation (11) holds for every stress point t_s, where t_s is now the accumulated net stress time. For the particular case of the RD model we will show later that this is an excellent approximation. The likely correctness of this assumption for real measurement data is demonstrated in the following. First, the application of the above procedure to the detailed relaxation data published by Reisinger *et al.* [7] is studied in Fig. 3 and for the IMEC data otherwise published in [5] in Fig. 4, where the universality is shown at three different temperatures, 50 °C, 125 °C, and 200 °C.

A. Uncertainty Related to the Analytic Expression for $r(\xi)$

In Fig. 3 and Fig. 4, the universality has been demonstrated using the relaxation expression (7). Naturally, one may inquire about the influence of the universal relaxation function on the final result. This issue is explored in Fig. 5, where the above procedure is performed using the universal relaxation functions (7), (8), and (9). Although all expressions considered here can be fit to the measurement data, the different predictions for the extrapolated value at $t_M \to 0$ result in an uncertainty of roughly 10%. Unfortunately, the data available to us at the moment is not conclusive to decide on the best approximation for r, but the power-law-like expression (7) gives the best fit over the entire range also including large relaxation times, while the log-like expression might slightly better capture the saturation for small times. The slight deviation for $\xi > 10^2$ could bear the hint of a permanent

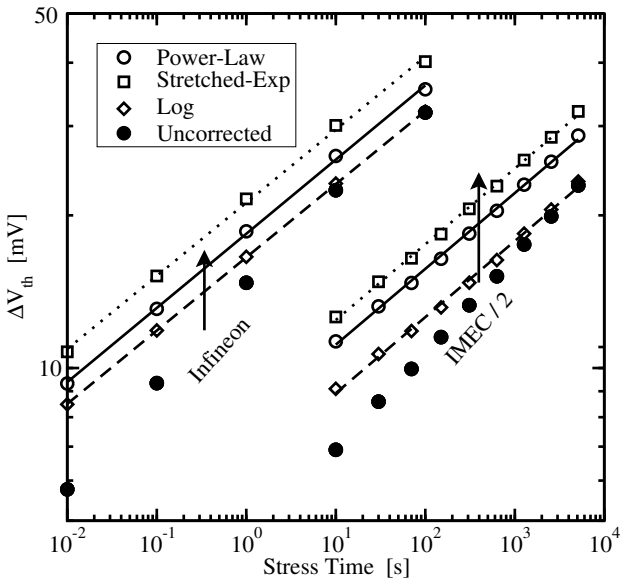

Fig. 6: Corrected results using the three universal relaxation functions for the MSM data of Reisinger *et al.* [7] (Infineon) and Kaczer *et al.* [5] (IMEC). Independently of the relaxation function the same power-law exponent is obtained while the prefactor depends much stronger on the choice of r. For the sake of clarity, the IMEC data has been divided by a factor of 2.

	Infineon		IMEC	
Analytic Form	A [mV]	n	A [mV]	n
Uncorrected	14.1	0.1866	8.9	0.1918
Log	16.5	0.1444	12.6	0.1499
Power-Law	18.4	0.1462	15.8	0.1495
Stretched-Exp	21.2	0.1441	17.6	0.1493

TABLE I: Influence of the various analytical expressions on the corrected slope and prefactor for a fit expression of the form At^n. Although the prefactor depends on the correction method, the more important slope is virtually independent of it.

component $P(t_s)$, which would have to be subtracted according to (4) before studying universality [27].

To judge the implications of the uncertainty related to the floating behavior of the final universal curve, we use the universal relaxation relation (11) to extrapolate from each measurement point to its 'true' value. Interestingly, one observes that in this particular case all relaxation functions result in the **same power-law exponent** while only the prefactors depend significantly on the choice of $r(\xi)$. This is demonstrated in Fig. 6 for the relaxation data of Reisinger *et al.* [7] and Kaczer *et al.* [5]. Interestingly, although the uncorrected data of Reisinger *et al.* have quite a visible curvature on a log-log plot and are as such not well described by a power-law, the corrected version can be very well described by a power-law with exponent $n = 0.15$. We recall that a deviation of the initial data from a power-law was one of the reasons why Reisinger *et al.* concluded that a log-like hole-trapping component has to be superimposed to a standard H-RD power-law with exponent $n = 1/4$. We thus conclude that even ultrafast measurements with $t_M = 1\,\mu$s **are not fast enough for obtaining the 'true' slope**, at least not for $t_s < 100$s. This can also be directly observed in Fig. 3 where only the data obtained for $t_s = 100$s shows signs of saturation.

The delayed and corrected values for the exponent and the prefactor are summarized in Table I, which confirms that the differences in

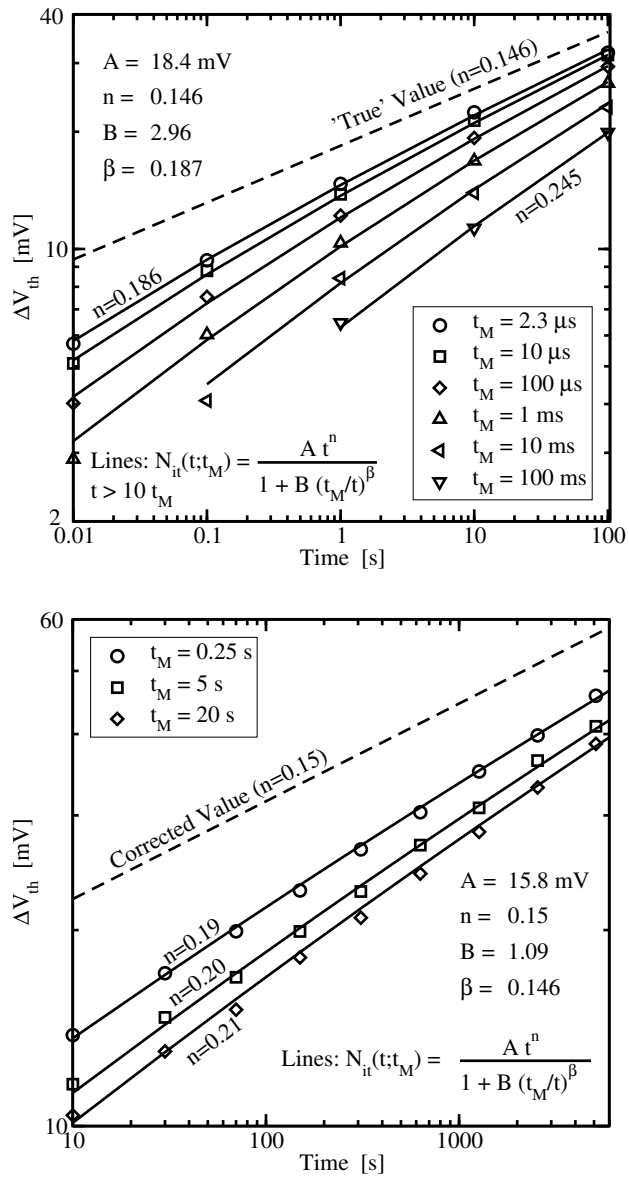

Fig. 7: Comparison of the analytic model for MSM measurements based on the universal relaxation to the data of Reisinger *et al.* [7] (top) and Kaczer *et al.* [5] (bottom). Excellent accuracy of the analytic model is obtained for all available delay times. In addition, the 'true' NBTI degradation can be recovered by extrapolating to $t_M = 0$s.

the values for the corrected exponents are well below the measurement accuracy. Thus, even if the empirical relaxation functions given by (7) – (9) contain some uncertainty, the influence on the corrected power-law exponent is probably small and the correct exponent can be expected to reflect the true degradation in a much better way than the uncorrected result. Also, the corrected exponents agree very well with exponents reported in some recent publications obtained from OTF measurements [9, 23].

B. Influence of Measurement-Delay on the Power-Law Parameters

Next, we show that the universal relaxation expression naturally connects individual stress curves obtained using the MSM technique with different delay times. For simplicity we assume that the true

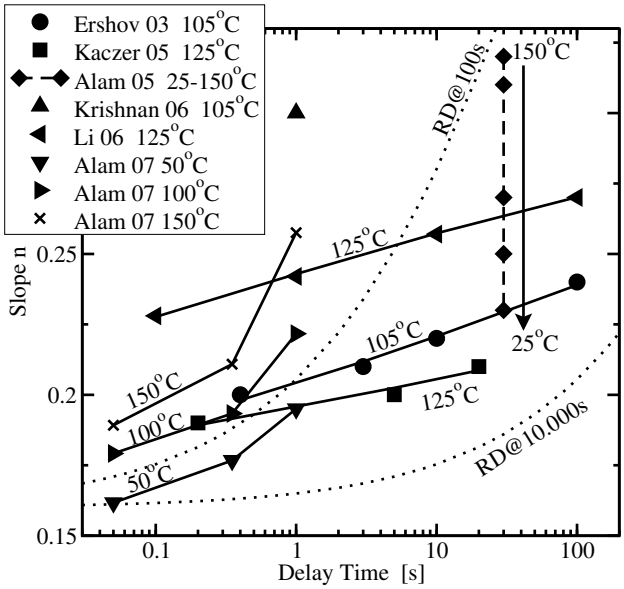

Fig. 8: Influence of the measurement delay on the measured slope as reported by various groups. The solid lines are given by a fit to (14) using the parameters in Table II. Note the strong temperature-dependence of the reported slopes and that the slopes were found to be constant over 3-4 decades in many measurements. Clearly, there is a large spread in the measurement data indicating a technology dependence. The dotted lines show the slopes predicted by the RD model at $t_s = 100\,s$ and $t_s = 10,000\,s$. Note that the RD slope changes considerably within two decades, is per construction temperature independent, and cannot be adjusted to the technology.

Source	T	n	B	β
Ershov [21]	105	0.15 (fixed)	1.49	0.179
Kaczer [5]	125	0.15 (fixed)	1.29	0.136
Li [33]	125	0.15 (fixed)	4.08	0.163
Alam [17]	50	0.155	4.79	0.611
Alam [17]	100	0.177	40.23	0.973
Alam [17]	150	0.186	102.2	1.048

TABLE II: The parameters for (14) used to fit the data in Fig. 8 assuming $t_s = 1,000\,s$. The fit was obtained using a fixed $n = 0.15$ with a simple least-square algorithm. However, in order to fit the data of [17], which are different from the other sources considered in this study, n had to be included as a free parameter. Keep in mind that these values should be taken with care, since they were extracted by a fit to three or four rather inaccurate slope values using two/three free parameters. The inaccuracy of the slope values is a result of both the measurement uncertainty as well as the time-dependence of the slope.

degradation behavior follows a power-law as $S(t_s) = At_s^n$ and that the universal relaxation is given by (7). Due to the measurement delay one observes instead of the power-law

$$S_M(t_s, t_M) = S(t_s)\, r(t_s, t_M) = \frac{At_s^n}{1 + B(t_M/t_s)^\beta} \ . \qquad (13)$$

Equation (13) is validated against the Infineon and IMEC data in Fig. 7 where the parameters B and β are given by the universal relaxation law. The analytic expression (13) exactly reproduces the delayed measurement results for various delay times t_M and thereby convincingly confirms our assumptions stated above.

As a consequence of the measurement delay, the *observed* power-law exponent n_M will be time-dependent and given through (13) as

$$n_M(t_s, t_M) = n - r'(t_M/t_s)\frac{t_M/t_s}{r} = n + \frac{\beta B}{B + (t_s/t_M)^\beta} \qquad (14)$$

with $r'(\xi) = \partial r(\xi)/\partial \xi$. It is worthwhile to stress that although many groups report a 'constant' measured power-law exponent over 3-4 decades which varies as a function of the temperature and delay-time, this can of course only be approximately correct. The fact that all curves obtained with different delay-times have to merge at larger times, makes a time-dependent slope a necessity. However, depending on the actual values of B and β this time-dependence will be more-or-less visible in a log-log plot. In general, the smaller β, the less visible the time-dependence will be. A comparison of measured power-law exponents as a function of the delay time t_M and temperature is given in Fig. 8. Most of the data shows an apparently constant power-law exponent (within the measurement accuracy) over 3-4 decades. Clearly, the measured power-law exponents, and consequently B and/or β (see Table II), depend on temperature and on the particular technology.

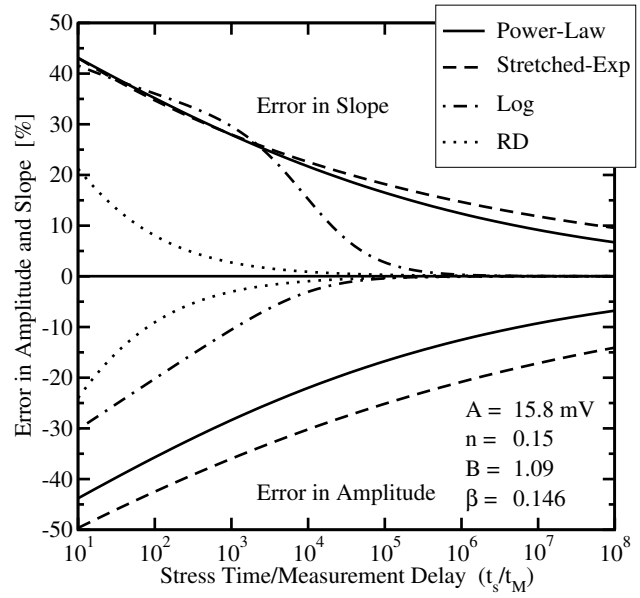

Fig. 9: Error in the slope (top) and prefactor (bottom) introduced by the measurement delay for the IMEC data of Fig. 4 as a function of t_s/t_M. Based on the power-law correction and extrapolation one obtains that for the slope and prefactor to be accurate be within 10%, the stress time has to be at least six orders of magnitude larger than the measurement delay. Also shown is the prediction of the RD model which reaches the 10% criterion after $t_s = 10 \times t_M$, in contradiction to measurements.

The measurement delay related error in the slope and prefactor of a supposed power-law degradation can be studied analytically using (13). This is shown in Fig. 9 where these errors are displayed as a function of t_s/t_M using the parameters extracted from the IMEC data of Fig. 4. Naturally, all models confirm that as a consequence of the delay the slope is always overestimated while the prefactor is always underestimated. Of particular interest here is the different prediction of the three universal relaxation expressions. The power-law and the stretched-exponential predict that in order to obtain the slope and prefactor with an accuracy of 10%, the stress time t_s has to be at least six orders of magnitude larger than the measurement delay t_M. For example, to keep the stress time below $10^4\,s$ requires a measurement delay smaller than 10 ms. In contrast, the logarithmic

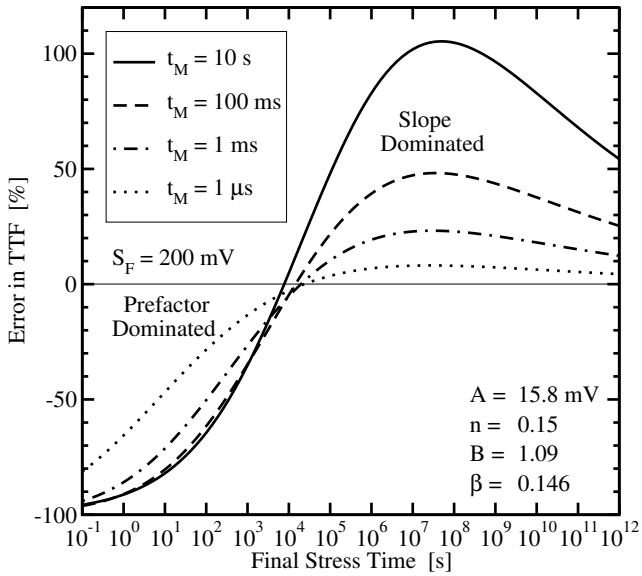

Fig. 10: The influence of the measurement delay and the stress time on the predicted time-to-failure. The exact shape and the intersection with the 0% line depends on the failure condition S_F. In general, however, the error in the prefactor dominates the result for short stress times while for large stress times the slope error prevails.

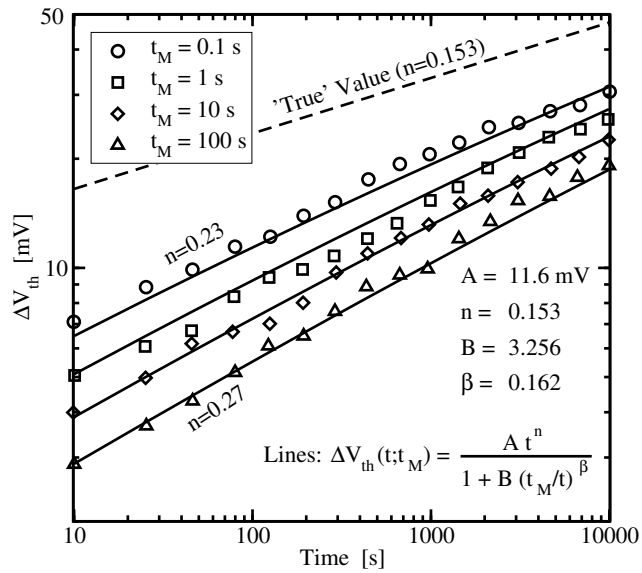

Fig. 11: Reconstruction of the 'true' degradation from MSM data obtained Li *et al.* [33] with four different delay times **without the knowledge of the detailed relaxation behavior**. Again, a corrected slope of about $n \approx 0.15$ is obtained. Note that even at $t_s = 10^4$ s the lines do not merge and the impact of the delay is still clearly visible.

relaxation expression shows a distinct kink and predicts the same accuracy after only four orders of magnitude. Further measurements are required to clarify which expression is more accurate. Also note that for $t_s < 10^3 \times t_M$ all models predict roughly the same error in the slope and consequently result in the same correction. This is in agreement with all measurement data known to us where the influence of the measurement delay is still clearly visible after long stress times. We finally remark that the RD model predicts an error smaller than 10% after $t_s = 10 \times t_M$. In other words, the RD model predicts that after a stress time not significantly longer than the measurement delay, the error introduced by the delay becomes negligible, which is in contradiction to measurements.

C. Influence of Measurement-Delay on Time-to-Failure Prediction

From the corrected power-law exponents and prefactors the time-to-failure can be straight-forwardly calculated. With the criterion $S(t_F) = S_F$ one obtains

$$t_F = \left(\frac{S_F}{A}\right)^{1/n} . \tag{15}$$

In contrast, an MSM measurement is normally conducted up to a certain stress time $t_s = t_0$ from which the exponent $n_M(t_0, t_M)$ and prefactor are approximately extracted and subsequently used for the extrapolation in (15). The error inherent in the MSM technique can be analytically estimated using (13). By considering a Taylor expansion around t_0 we obtain $S_M(t_s, t_M) \approx S_M(t_0, t_M)(t/t_0)^{n_M(t_0,t_M)}$ and consequently

$$\tilde{t}_F(t_0, t_M) = t_0 \left(\frac{S_F}{S_M(t_0, t_M)}\right)^{1/n_M(t_0,t_M)} . \tag{16}$$

The overall error in this time-to-failure approximation consists of a contribution from the error in the prefactor and the error in the slope and the exact behavior of $\tilde{t}_F(t_M)$ thus depends on t_0, t_M, and S_F. A typical scenario is shown in Fig. 10 with the parameters from Fig. 7. As already pointed out by Schlünder *et al.* [34], the time-to-failure is first underestimated and then overestimated. The underestimation is a consequence of the underestimated prefactor which dominates at early stress times but is overcompensated by the overestimated slope at large stress times. The intersection with the 0% line depends on the failure condition S_F. In particular, for delay times larger than 1 s the error can be quite significant, independent of the final stress time t_0, meaning that **the detrimental effect of delay does not disappear, even at very large stress times**.

D. Reconstruction of MSM Data without Detailed Relaxation Data

We summarize the particularly intriguing feature of the universal relaxation in the context of MSM data which is given by the fact that it **allows one to reconstruct the 'true' (undelayed) measurement curve** from delayed data sets. This suggests a **novel measurement technique**: (i) determine $R(t_s, t_M)$ using different delay times, or, determine $R(t_s, t_M)$ using a single delay time and add a long relaxation period at the end. (ii) from that data determine B and β. (iii) Finally, calculate the 'true' degradation using (11). The variant where B and β have been obtained from relaxation data has been already demonstrated in Fig. 7. However, the method also works for MSM data obtained with different delay times where no relaxation data is available. In that case the parameters A, n, B, and β can be directly extracted through fitting of equation (13). This is demonstrated in Fig. 11 for the data published by Li *et al.* [33]. Again, the extracted parameter values agree very well with the cases where we had access to the full relaxation data.

V. The Reaction-Diffusion Model

Based on the universal relaxation we now analyze the predictions obtained from theoretical models. For this, the models under consideration have been implemented into a partial-differential-equation solver and solved numerically, to rule out any uncertainties related to approximate analytic expressions. We start with the reaction-diffusion

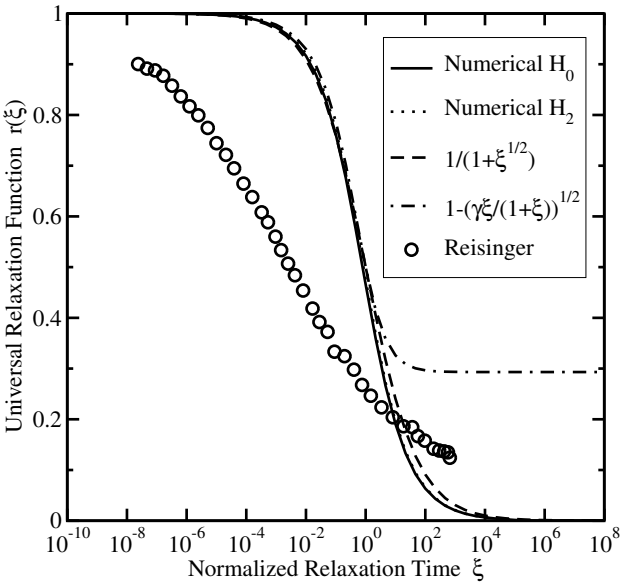

Fig. 12: Comparison of the two analytic expressions for the RD relaxation behavior with numerical results obtained for H and H_2 kinetics. The power-law-like expression is accurate for all relaxation times and will be used as reference throughout this work. Due to the lack of parameters there is **no way** to fit the measurement data with the RD model.

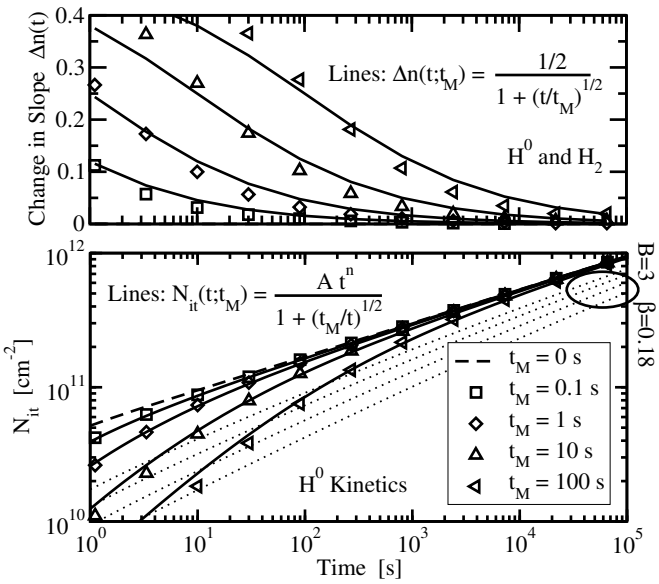

Fig. 13: Influence of the measurement delay t_M as predicted by the RD model. Comparison of the analytic model (lines) with the numerical solution (symbols) proves the excellent accuracy of the analytic model for $t > t_M$. Note that the RD model predicts a very small influence of delay for longer stress times, in contrast to Fig. 7. For the sake of comparison, a more realistic influence of the measurement delay is given by the dotted lines, obtained with typical parameter values $B = 3$ and $\beta = 0.18$.

(RD) model, which predicts a universal relaxation practically independent of the species (H and H_2) as $r(\xi) = 1/(1+\xi^{1/2})$. This analytic expression is compared to the numerical results for both species in Fig. 12. Also shown is the measurement data of Reisinger *et al.* [7]. It is worthwhile to realize that the relaxation predicted by the RD model **does not depend on any model parameters**. The reason for this is that just as during the stress phase, relaxation occurs in the diffusion-limited regime. However, since the forward term is dropped and the interface state density $\Delta N_{it}(t)$ is assumed to be in quasi-equilibrium, the actual value of the reverse rate has no influence on the result. Also, since we have two diffusion fronts, one diffusing towards the interface to passivate interface defects and one diffusing away from it, the diffusion coefficient itself cancels out of the expression. Consequently, it must be clearly emphasized that since the relaxation predicted by the RD model cannot be made to depend on gate bias, temperature and process conditions, there is no way to fit that particular measurement data. In particular, $\beta = 1/2$ is much larger than observed experimentally, leading to a relaxation which is too slow in the beginning and too fast in the end. This is clearly visible in Fig. 12 where most of the relaxation occurs within 3 decades whereas the measurements show relaxation over 10 decades. Consequences of this erroneous relaxation prediction are a **heavily time-dependent but temperature-*in*dependent slope** in the RD simulated delayed measurements, and a vanishing influence of the delay on the measurement result for $t_s \gtrsim 10 \times t_M$ (Fig. 13), in contradiction to measurements [5, 21], see also [19]. The only way to move the relaxation curve to shorter relaxation times is to bring the forward reaction into the quasi saturation regime where hydrogen has already piled up considerably in the oxide (assuming a reflecting boundary condition). However, in addition to the fact that this behavior is not universal, the slope during the stress phase approaches zero.

VI. Extended Classical Reaction-Diffusion Models

As the standard form of the RD model has been found to have also limitations during the stress phase [9, 35], extended versions have been introduced. However, the question of whether these extended models are better able to describe the relaxation behavior has so far only been qualitatively assessed and a rigorous statement is missing. This will be done in the following.

A. Two-Region RD Model

First, it has been noted that due to the extremely thin oxides used in modern technology, the diffusing hydrogen species may quickly reach the oxide/poly interface [9]. As a consequence, the degradation will be dominated by the presumably slower diffusion in the poly gate. We will discuss two variants of RD models extended to account for such a situation. The first variant assumes the oxide/poly interface to be a perfect transmitter. At short times the oxide will be filled with H_2. At later times, the overall hydrogen diffusion is dominated by the slower diffusion inside the poly-gate and the model behaves just like the standard H_2-RD model. One might suspect that the hydrogen stored inside the oxide, where we have assumed the diffusivity to be larger, modifies the relaxation behavior. Under certain conditions this is indeed the case, with undesired properties, though, as shown in Fig. 14. For large stress times, most hydrogen is stored in the poly and the model predicts the same relaxation as the RD model. Thus, in order to see the influence of the two regions we have to look at shorter stress times, in our particular case $t_s = 10$ s and $t_s = 100$ s, where the population in both regions is of the same order of magnitude. However, despite the fact that the shape of the relaxation curve does not agree with measurement data, since the relaxation is determined by both reservoirs, it depends on the ratio of the reservoir

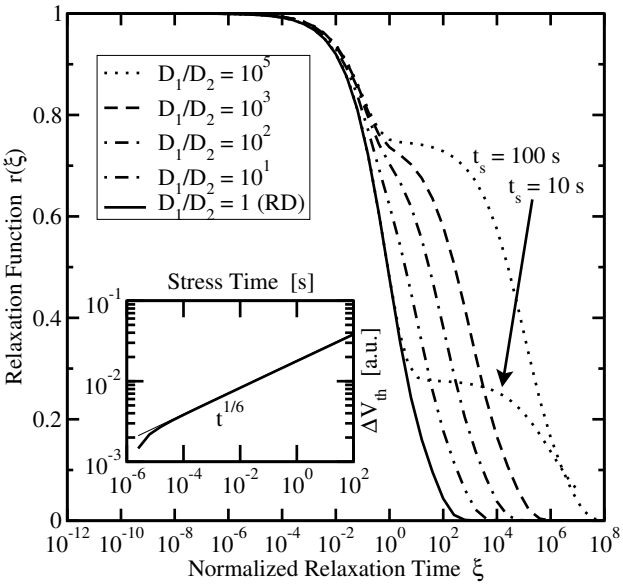

Fig. 14: Numerical simulation of a generalized RD model with two different diffusion coefficients in the oxide and poly. For this particular set of parameters, no difference is visible during the stress phase, while the relaxation behavior slows down and displays non-universal humps.

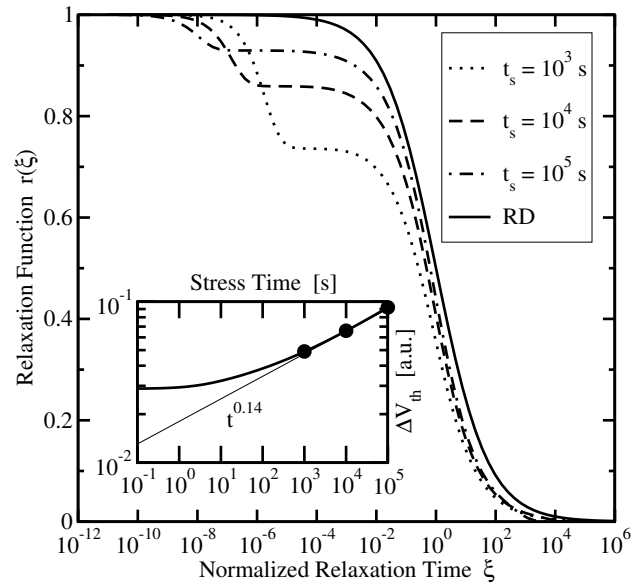

Fig. 15: Numerical simulation of the two-interface RD model relaxation behavior. Although faster relaxation than with the standard RD model is possible, the relaxation is not universal. In order to obtain a visible influence on the relaxation behavior, the hydrogen stored in both regions has to be of the same order of magnitude, which has a detrimental influence on the power-law slope during the stress phase.

occupancies, which changes with time and consequently results in a non-universal relaxation as is clearly visible in Fig. 14.

B. Two-Interface RD Model

The two-interface model assumes that atomic hydrogen is released from the silicon/oxide interface which then diffuses through the thin oxide and depassivates defects at the oxide/poly interface [9, 36]. The creation of defects at the opposite interface is supported by SILC measurements [9]. The hydrogen from the oxide and the released hydrogen at the oxide/poly interface diffuse as H_2 through the poly and result in an overall power-law exponent of $1/6$ at large times. It has been reported that such a two-interface model may predict a faster recovery compared to the standard RD model [36]. For this to be the case, the amount of 'fast' hydrogen stored in the oxide must be of the same order of magnitude compared to the 'slow' hydrogen stored in the poly. As in the case of the two-region RD model, it is again possible to modify the relaxation behavior to a certain extent, see Fig. 15. In this case the relaxation can be made faster than with the standard RD model because the 'fast' hydrogen concentration inside the oxide is saturated, resulting in a shift to smaller normalized relaxation times ξ on the universal plot. However, just as with the two-region model, the resulting relaxation is not universal, as the ratio of these two hydrogen storage areas changes with time, see Fig. 15.

C. Explicit H-H_2 *Conversion RD Model*

It has been reported that measurements display a power-law exponent of $1/3$ during the initial stress phase [8, 17]. This has been explained by an extended RD model which explicitly accounts for the dimerization of H into H_2 [17], rather than assuming an instantaneous transition, in addition to the diffusion of both hydrogen species.

$$\frac{\partial[\mathrm{H}]}{\partial t} = k_{\mathrm{H}_2}[\mathrm{H}_2] - k_{\mathrm{H}}[\mathrm{H}]^2 \tag{17}$$

Depending on the values of k_{H_2} and k_{H}, either pure H or H_2 kinetics can be observed. In addition, a regime with the aforementioned

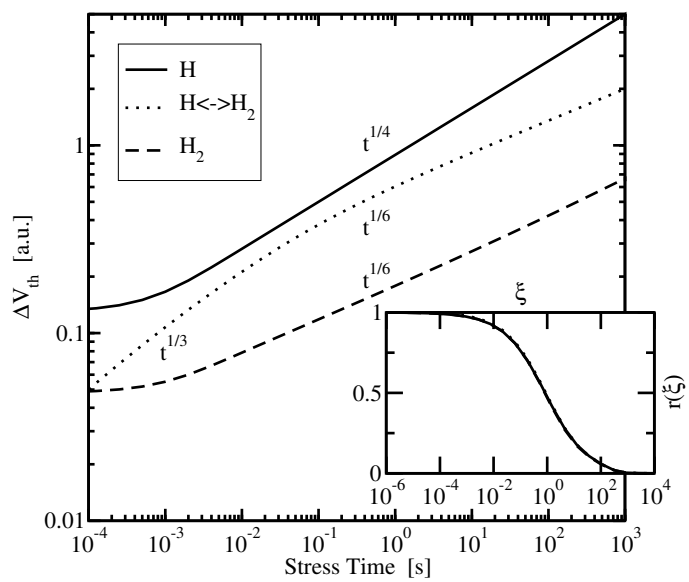

Fig. 16: Numerical simulation of the generalized RD model with explicit H to H_2 conversion. During the stress phase, the model gives power-law exponents known from the H and the H_2 models, in addition to a transitional region with $n = 1/3$, while the relaxation behavior cannot be influenced by any of the available parameters.

transitional power-law exponent of $1/3$, which eventually changes to $1/6$, is possible. Since measurement data suggests a long-term exponent closer to $1/6$ than to $1/4$, the parameters have to be chosen in such a way that the total amount of stored $[\mathrm{H}_2]$ is much larger than $[\mathrm{H}]$. One might conclude from this that the two distinct reservoirs of H and H_2 may allow for a modified relaxation behavior. However,

this is not the case for the simple reason that the model stays within the limits set by pure H and H_2 behavior, just as during the stress phase. Since the relaxation of both species is practically equivalent, **no influence on the relaxation behavior** is obtained from such a model, see Fig. 16.

D. Final Notes On Reaction-Diffusion Models

We have shown that irrespective of the extensions applied to the RD model, the recovery behavior cannot be described with the published models in their present form. The fact that some OTF measurements and the corrected 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 view point that the RD model correctly covers the stress part while 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 reaction [17]. Without the backward reaction, which inserts 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 dropped, 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 and there constructively changing the reaction-limited exponent of $n = 1$ to the 'correct' diffusion-limited value of $n = 1/6$.

VII. Dispersive NBTI Models

It has been clearly shown in the previous sections that the RD model predicts 80% of the relaxation to occur within 3 decades, while in reality relaxation is observed to span more than 10 decades [5–7]. **This indicates some form of dispersion in the underlying physical mechanism(s)**. Various forms of dispersion have already been introduced into NBTI models based on either (i) diffusion [3–5], (ii) hole tunneling from/into states in the oxide [16], and (iii) reaction rates at the interface [2, 37]. The models suggested to capture these mechanisms will be benchmarked in the following using the universality as a metric.

A. Reaction-Dispersive-Diffusion Models

First, we consider generalized RD models based on **dispersive transport** of the hydrogen species [5, 38]. Interestingly, published results based on dispersive transport equations have demonstrated both an increase [15, 36, 39] as well as a decrease [5, 40, 41] of the power-law exponent with increasing dispersion. This discrepancy has been identified to be a consequence of the different boundary conditions employed in the reverse reaction at the interface [38]. Since dispersive transport is a consequence of most of the hydrogen being trapped in a broad distribution of traps, exponential in many models, one might either consider the *total* hydrogen concentration ($H_{it} = H_{tot}$), that is include both mobile and trapped hydrogen to result in a decreased exponent, or, just allow the mobile hydrogen ($H_{it} = H_c$) to passivate dangling bonds, which results in an increased exponent. A previous analysis of the dispersive transport equations [5] was based on various assumptions (such as pulse-like excitation [38], uncertainties in the boundary conditions [38], and a neglected history of previously trapped hydrogen atoms during relaxation) which led to only approximative solutions. As it turned out, a more rigorous derivation is rather involved. An approximation for $\xi < 1$ ($t_r < t_s$, as normally encountered during measurements), is given by (4) with

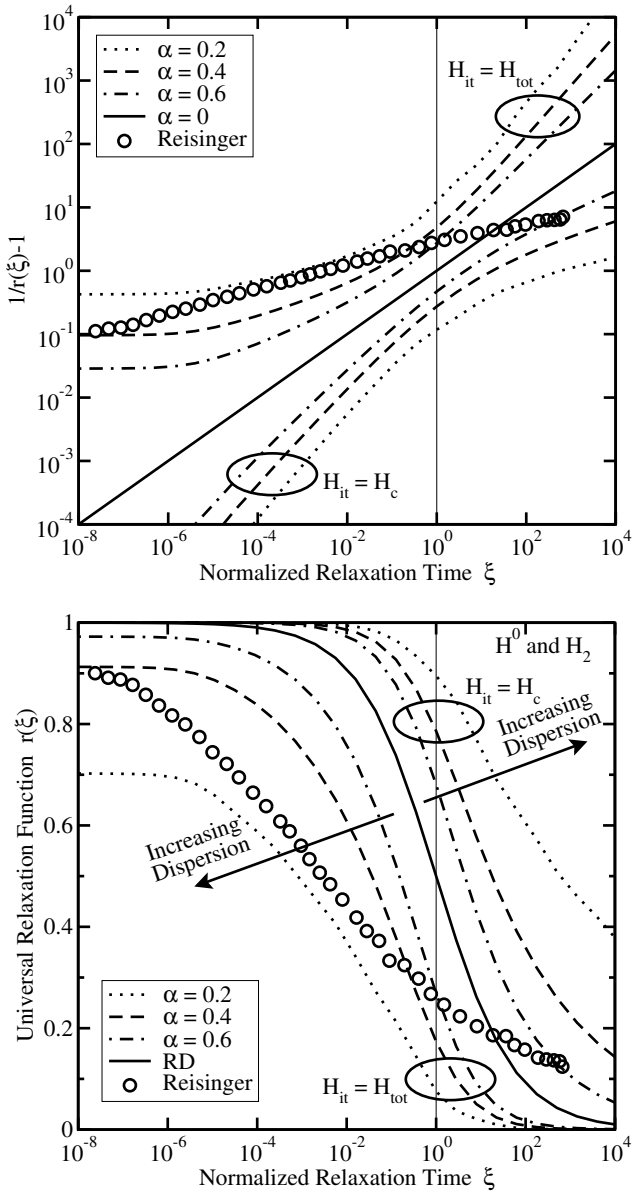

Fig. 17: Universal relaxation as predicted by the full numerical solution of the dispersive transport models for various values of the dispersion parameter α. Also shown is the data from Reisinger as a reference. The model with $H_{it} = H_{tot}$ always stays below the diffusive (RD) limit, while the model with $H_{it} = H_c$ always stays above. The diffusive limit is like a watershed which cannot be crossed by either model. Also note that the $H_{it} = H_{tot}$ model appears to have a limit different from unity for $\xi = 0$, which is a result of an extremely fast relaxation triggered by the hydrogen stored right at the interface.

B and β depending on the boundary condition and the dispersion coefficient α. Interestingly, for $\xi > 1$ the behavior changes and different values for B and β have to be used (cf. Fig. 17).

In order to avoid any uncertainties inherent in approximate analytical solutions, we numerically solve the full time-dependent multiple trapping model [42] to allow for an accurate description of both the stress and the relaxation phase. The results shown in Fig. 17 display a much broader range of possible relaxation characteristics compared to classic diffusion. Nevertheless, the dispersive transport models in their present form are not able to fully explain the experimentally

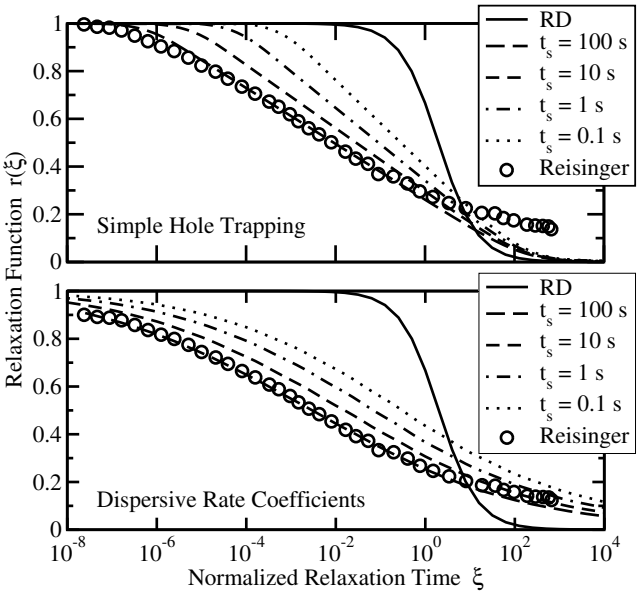

Fig. 18: Top: The simple dispersive hole trapping model used in [16] can be fit to an individual relaxation curve (shown is the 'log' variant of Fig. 5) but does not scale universally and gives a very small slope during the stress phase. Bottom: The dispersive rate model can be used to fit the data for an individual relaxation curve, the power-law variant of Fig. 5 at $t_s = 100\,\mathrm{s}$ in this case, but does not scale universally either. In addition, the calibrated model gives a rather strong curvature during the stress phase with a too small power-law exponent $n \approx 0.03$, the variance of the rate coefficients had to be set to a value considerably larger than reported ($\sigma_A = 0.211\,\mathrm{eV}$ and $\sigma_D = 0.264\,\mathrm{eV}$, compare [43, 44]), and in general the model cannot be fit to both the stress and relaxation phase.

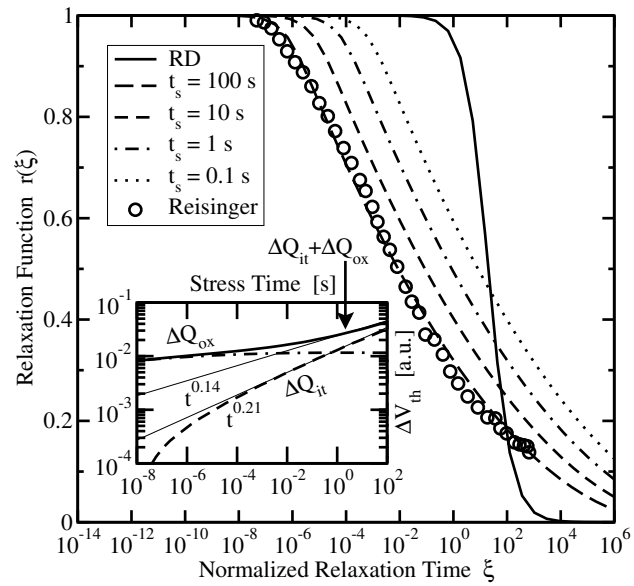

Fig. 19: Relaxation as predicted by the Tewksbury model. The model can be fit to the data for an individual relaxation curve, shown is the 'log' variant at 100 s, but does not scale universally. Also, the excellent fit comes at the price of a very small power-law exponent at early times during the stress phase, in contradiction to Fig. 4 of [7]. It is not possible to fit both the stress and relaxation phases at the same time.

observed relaxation. Also note that the $H_{it} = H_{tot}$ model appears to have a limit different from unity for $r(\xi \to 0)$. This is a result of the extremely fast relaxation triggered by the hydrogen stored directly at the interface. The exact shape of this initial hump (not shown) depends on the stress time and the width of the interfacial layer, thereby rendering this model non-universal.

B. Dispersive-Rate Coefficients

Next we consider reaction-limited models using a **dispersion in the rate coefficients** [2, 37]. Huard *et al.* [2] suggested that the forward rate is dispersive and that the generated interface states do not relax at all, or at least not at shorter relaxation times [27]. A numerical solution of a generalized model which considers both a dispersive forward and reverse reaction [43] results in an apparently very flexible model which can be nicely fit to a single relaxation curve. Unfortunately, however, the excellent fit during the single relaxation phase adversely affects the quasi-power law exponent during the stress phase which reduces to very small values ($n \approx 0.03$). In addition, the model does not scale universally. This is demonstrated in Fig. 18.

C. Dispersive Hole Trapping Models

In addition to the creation of interface states, trapped charges have been made responsible for the observed threshold voltage shift during NBT stress. In particular it has been argued that these charges are responsible for the fast component observed both during stress and relaxation [2, 16]. A simple phenomenological hole trapping model has been put forward by Yang *et al.* [16] based on a broad distribution of trapping times. Again, although a single relaxation curve can be nicely fit, the model does not scale universally either, see Fig. 18.

A more detailed hole trapping model has been derived by Tewksbury [32] based on the various possible transitions from conduction, valence, and interface states into bulk oxide traps. Its use for NBTI has been suggested by Huard *et al.* [18] to cover the recoverable part of the degradation. For the following discussion we limit ourselves to the component of the model which results from charge transfer from an interface state into an oxide trap and back to the interface state, the other suggested mechanisms behave similarly and follow analogously. During stress, the trapped charge accumulated via transfer from the interface states can be given as $S(t_s) \approx A \ln(t_s/\tau_{0s})$ while the *absolute* relaxation is given by $R(t_s, t_r) \approx A \ln(t_s \tau_{0r}/(t_r \tau_{0s}))$. The latter is an approximate form of (10) and considers two different time-constants for capture and emission. Using the previous two relations, the relaxation function is given by

$$r(t_s, t_r) \approx 1 - \ln\left(\frac{t_r}{\tau_{0r}}\right) \ln^{-1}\left(\frac{t_s}{\tau_{0s}}\right) , \tag{18}$$

which cannot be written as a function of t_s/t_r and is consequently not universal. The full numerical solution of the Tewksbury model is given in Fig. 19 together with an excellent fit for a single relaxation curve. However, in order to obtain such a fit, the logarithmic behavior of the hole trapping component results in a slope close to zero during the stress phase. Also shown in Fig. 19 is a permanent component modeled by a numerical solution of a dispersive forward rate only, as suggested by Huard *et al.* [18]. Note however, that after a certain stress time the damage will be dominated by ΔQ_{it} and the observed relaxation given only through ΔQ_{ox} will be minimal. This is also not compatible with the data at hand where even at large stress times considerable relaxation can be observed.

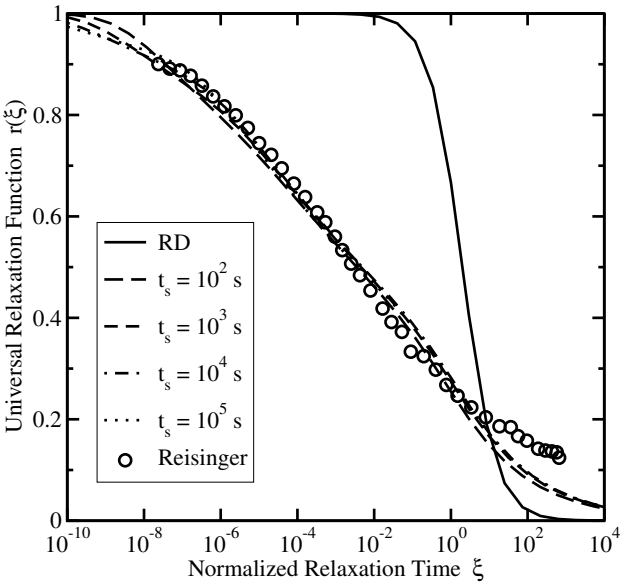

Fig. 20: Best universal scaling was obtained by combining dispersive hydrogen transport with dispersive hole transport [45]. Universality is obtained when the hole trap density is determined by the trapped hydrogen which is assumed to create hole-traps.

D. Multiple Mechanisms

Since none of the studied mechanisms can fully capture the universal relaxation, we have considered various combinations in our numerical framework. In order to obtain a universal behavior, some points need to be considered. Consider the case that the total observed threshold voltage shift is the result of two independent mechanisms, that is, $S = S_1 + S_2$. During relaxation one observes $R = S_1 r_1 + S_2 r_2$ and the normalized relaxation function is given by $r = \rho r_1 + (1-\rho) r_2$ with $\rho = S_1/(S_1+S_2)$. If the two degradation mechanisms progress differently with time, ρ will be a function of t_s and r cannot be universal. We thus conclude that for the relaxation to be universal, the two mechanisms need to be tightly coupled, that is, $S_1/S_2 = \text{const}$, or at least roughly constant within the range of measured ξ and within the measurement tolerance. Another option is that both mechanisms relax equally, $r_1 \approx r_2$.

Promising results have been obtained by **combining dispersive hydrogen transport** with a **dispersive hole transport model** [45] (cf. Fig. 20). This combination is consistent with qualitative statements found in literature [2, 9, 16]. In order to make the fast hole trapping component universal, the hole trap density was determined from the trapped hydrogen density in a self-consistent manner. Nevertheless, this is just a first attempt and the microscopic details of such a model need to be considered in greater detail.

VIII. Conclusions

We have thoroughly analyzed the relaxation of NBT stress induced damage using data from various groups. The observed universal relaxation behavior has been quantified and modeled using possible empirical expressions. It has been demonstrated that data obtained via conventional MSM sequences can be analytically described as a function of the delay introduced during the measurement. In particular, this analytic expression allows one to **reconstruct a corrected 'true' degradation curve**. Using this corrected curve, it might be possible to more accurately estimate the time-to-failure.

We have then used the relaxation behavior and in particular the universality as a benchmark for existing NBTI models. There we have found that none of the existing models is capable of reproducing both the stress and the relaxation phase with the same set of parameters. While the classic reaction-diffusion model scales universally, it predicts relaxation to occur mainly during three decades, in contradiction to detailed relaxation measurements available in literature which span more than ten decades. No improvement could be found in extended RD models using two regions, a second interface, or an explicit transition from atomic to molecular hydrogen. Models based on an extension of the reaction-diffusion model with dispersive transport, somewhat improve on the situation but are still not able to cover the whole relaxation regime. Other dispersive models, like dispersive forward and backward rates or dispersive hole-trapping models allow one to fit an individual relaxation curve only but are not universal. In addition, we were not able to describe both the stress and relaxation phase with the same set of parameters. This indicates a significant gap in our current understanding of NBTI.

We particularly wish to point out that it is of utmost importance not to consider the inaccuracies of existing models during the relaxation phase of secondary importance compared to the stress phase. The reason for this is only partially related to the frequently quoted fact that continuous DC stress is rarely observed in a circuit and that duty-cycle dependent corrections have to be applied. The more important point we want to make here is that during the stress phase the relaxation mechanism in existing models always interacts with the degradation mechanism, forming the overall time behavior during the stress phase. It is only during the relaxation phase, where the degradation mechanism is absent, that the relaxation mechanism can be studied in full detail, despite the difficulties arising during measurements. We therefore stipulate that a more complete NBTI model needs to **focus on the relaxation phase first** before attempting to cover the stress phase as well.

References

[1] M.A. Alam and S. Mahapatra, "A Comprehensive Model of PMOS NBTI Degradation," *Microelectr.Reliab.*, vol. 45, no. 1, pp. 71–81, 2005.

[2] V. Huard, M. Denais, F. Perrier, N. Revil, C. Parthasarathy, A. Bravaix, and E. Vincent, "A Thorough Investigation of MOSFETs NBTI Degradation," *Microelectr.Reliab.*, vol. 45, no. 1, pp. 83–98, 2005.

[3] M. Houssa, "Modelling Negative Bias Temperature Instabilities in Advanced p-MOSFETs," *Microelectr.Reliab.*, vol. 45, no. 1, pp. 3–12, 2005.

[4] S. Zafar, "Statistical Mechanics Based Model for Negative Bias Temperature Instability Induced Degradation," *J.Appl.Phys.*, vol. 97, no. 10, pp. 1–9, 2005.

[5] 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. IRPS*, 2005, pp. 381–387.

[6] 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. IEDM*, 2004, pp. 109–112.

[7] 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. IRPS*, 2006, pp. 448–453.

[8] S. Rangan, N. Mielke, and E.C.C. Yeh, "Universal Recovery Behavior of Negative Bias Temperature Instability," in *Proc. IEDM*, 2003, pp. 341–344.

[9] A.T. Krishnan, C. Chancellor, S. Chakravarthi, P.E. Nicollian, V. Reddy, A. Varghese, R.B. Khamankar, and S. Krishnan, "Material Dependence of Hydrogen Diffusion: Implications for NBTI Degradation," in *Proc. IEDM*, 2005, pp. 688–691.

[10] 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. IRPS*, 2006, pp. 735–736.

[11] Y. Miura and Y. Matukura, "Investigation of Silicon-Silicon Dioxide Interface Using MOS Structure," *Jap.J.Appl.Phys.*, vol. 5, pp. 180, 1966.

[12] D.K. Schroder and J.A. Babcock, "Negative Bias Temperature Instability: Road to Cross in Deep Submicron Silicon Semiconductor Manufacturing," *J.Appl.Phys.*, vol. 94, no. 1, pp. 1–18, 2003.

[13] D.K. Schroder, "Negative Bias Temperature Instability: What Do We Understand?," *Microelectr.Reliab.*, 2006, (online version).

[14] J.P. Campbell, P.M. Lenahan, A.T. Krishnan, and S. Krishnan, "Direct Observation of the Structure of Defect Centers Involved in the Negative Bias Temperature Instability," *Appl.Phys.Lett.*, vol. 87, no. 20, pp. 1–3, 2005.

[15] T. Grasser, R. Entner, O. Triebl, H. Enichlmair, and R. Minixhofer, "TCAD Modeling of Negative Bias Temperature Instability," in *Proc. SISPAD*, Monterey, USA, Sept. 2006, pp. 330–333.

[16] T. Yang, C. Shen, M.-F. Li, C.H. Ang, C.X. Zhu, Y.-C. Yeo, G. Samudra, S.C. Rustagi, M.B. Yu, and D.-L. Kwong, "Fast DNBTI Components in p-MOSFET with SiON Dielectric," *IEEE Electron Device Lett.*, vol. 26, no. 11, pp. 826–828, 2005.

[17] M.A. Alam, H. Kufluoglu, D. Varghese, and S. Mahapatra, "A Comprehensive Model for PMOS NBTI Degradation: Recent Progress," *Microelectr.Reliab.*, 2006, (online version).

[18] V. Huard, C.R. Parthasarathy, C. Guerin, and M. Denais, "Physical Modeling of Negative Bias Temperature Instabilities for Predictive Extrapolation," in *Proc. IRPS*, 2006, pp. 733–734.

[19] C. Shen, M.-F. Li, C. E. Foo, T. Yang, D.M. Huang, A. Yap, G.S. Samudra, and Y.-C. Yeo, "Characterization and Physical Origin of Fast V_{th} Transient in NBTI of pMOSFETs with SiON Dielectric," in *Proc. IEDM*, 2006, pp. 333–336.

[20] M. Denais, V. Huard, C. Parthasarathy, G. Ribes, F. Perrier, N. Revil, and A. Bravaix, "Interface Trap Generation and Hole Trapping under NBTI and PBTI in Advanced CMOS Technology with a 2-nm Gate Oxide," *IEEE Trans.Device and Materials Reliability*, vol. 4, no. 4, pp. 715–722, 2004.

[21] M. Ershov, R. Lindley, S. Saxena, A. Shibkov, S. Minehane, J. Babcock, S. Winters, H. Karbasi, T.Yamashita, P. Clifton, and M. Redford, "Transient Effects and Characterization Methodology of Negative Bias Temperature Instability in pMOS Transistors," in *Proc. IRPS*, 2003, pp. 606–607.

[22] V. Huard, M. Denais, and C. Parthasarathy, "NBTI Degradation: From Physical Mechanisms to Modelling," *Microelectr.Reliab.*, vol. 46, no. 1, pp. 1–23, 2006.

[23] 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. IEDM*, Dec. 2005, pp. 1–4.

[24] M.-F Li, G. Chen, C. Shen, X.P. Wang, H.Y. Yu, Y.-C. Yeo, and D.L. Kwong, "Dynamic Bias-Temperature Instability in Ultrathin SiO_2 and HfO_2 Metal-Oxide-Semiconductor Field Effect Transistors and Its Impact on Device Lifetime," *Jap.J.Appl.Phys.*, vol. 43, no. 11B, pp. 7807–7814, 2004.

[25] C.R. Parthasarathy, M. Denais, V. Huard, G. Ribes, E. Vincent, and A.Bravaix, "New Insights into Recovery Characteristics Post NBTI Stress," in *Proc. IRPS*, 2006, pp. 471–477.

[26] J.P. Campbell, P.M. Lenahan, A.T. Krishnan, and S. Krishnan, "Observations of NBTI-Induced Atomic-Scale Defects," *IEEE Trans.Device and Materials Reliability*, vol. 6, no. 2, pp. 117–122, 2006.

[27] V. Huard, "Private Communications".

[28] S. Tsujikawa, T. Mine, K. Watanabe, Y. Shimamoto, R. Tsuchiya, K. Ohnishi, T. Onai, J. Yugami, and S. Kimura, "Negative Bias Temperature Instability of pMOSFETs with Ultra-Thin SiON Gate Dielectrics," in *Proc. IRPS*, 2003, pp. 183–188.

[29] M.A. Alam, "A Critical Examination of the Mechanics of Dynamic NBTI for PMOSFETs," in *Proc. IEDM*, 2003, pp. 345–348.

[30] S. Tsujikawa and J. Yugami, "Positive Charge Generation Due to Species of Hydrogen During NBTI Phenomenon in pMOSFETs With Ultra-Thin SiON Gate Dielectrics," *Microelectr.Reliab.*, vol. 45, no. 1, pp. 65–69, 2005.

[31] J. Kakalios, R. A. Street, and W. B. Jackson, "Stretched-Exponential Relaxation Arising from Dispersive Diffusion of Hydrogen in Amorphous Silicon," *Physical Review Letters*, vol. 59, no. 9, pp. 1037–1040, 1987.

[32] T.L. Tewksbury and H.-S. Lee, "Characterization, Modeling, and Minimization of Transient Threshold Voltage Shifts in MOSFETs," *IEEE J.Solid-State Circuits*, vol. 29, no. 3, pp. 239–252, 1994.

[33] J.-S. Li, M.-G. Chen, T. P.-C. Juan, and K. Su, "Effects of Delay Time and AC Factors on Negative Bias Temperature Instability of PMOSFETs," in *IIRW Final Rep.*, 2006, pp. 16–19.

[34] C. Schlünder, W. Heinrigs, W. Gustin, and H. Reisinger, "On the Impact of NBTI Recovery Phenomena on Lifetime Prediction of modern p-MOSFETs," in *IIRW Final Rep.*, 2006, pp. 1–4.

[35] M.A. Alam, "NBTI: A Simple View of a Complex Phenomena," in *Proc. IRPS*, 2006, (Tutorial).

[36] S., Chakravarthi, A.T. Krishnan, V. Reddy, and S. Krishnan, "Probing Negative Bias Temperature Instability Using a Continuum Numerical Framework: Physics to Real World Operation," *Microelectr.Reliab.*, 2006, (online version).

[37] M. Houssa, M. Aoulaiche, J.L. Autran, C. Parthasarathy, N. Revil, and E. Vincent, "Modeling Negative Bias Temperature Instabilities in Hole Channel Metal-Oxide-Semiconductor Field Effect Transistors with Ultrathin Gate Oxide Layers," *J.Appl.Phys.*, vol. 95, no. 5, pp. 2786–2791, 2004.

[38] T. Grasser, W. Gös, and B. Kaczer, "Modeling of Dispersive Transport in the Context of Negative Bias Temperature Instability," in *IIRW Final Rep.*, 2006, pp. 5–10.

[39] 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. IRPS*, 2004, pp. 273–282.

[40] S. Zafar, M. Yang, E. Gusev, A. Callegari, J. Stathis, T. Ning, R. Jammy, and M. Ieong, "A Comparative Study of NBTI as a Function of Si Substrate Orientation and Gate Dielectrics (SiON and SiON/HfO_2)," in *Proc. VLSI Symp.*, 2005, pp. 128–129.

[41] M. Houssa, M. Aoulaiche, S. De Gendt, G. Groeseneken, M.M. Heyns, and A. Stesmans, "Reaction-Dispersive Proton Transport Model for Negative Bias Temperature Instabilities," *Appl.Phys.Lett.*, vol. 86, no. 9, pp. 1–3, 2005.

[42] J. Noolandi, "Multiple-Trapping Model of Anomalous Transit-Time Dispersion in *a*-Se," *Phys.Rev.B*, vol. 16, no. 10, pp. 4466–4473, 1977.

[43] 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.

[44] 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.

[45] J. Noolandi, "Equivalence of Multiple-Trapping Model and Time-Dependent Random Walk," *Phys.Rev.B*, vol. 16, no. 10, pp. 4474–4479, 1977.