Simultaneous Extraction of Recoverable and Permanent Components Contributing to Bias-Temperature Instability

T. Grasser*, B. Kaczer°, P. Hehenberger*, W. Gös*, R. O'Connor°, H. Reisinger†, W. Gustin†, and C. Schlünder†

* Christian Doppler Laboratory for TCAD at the Institute for Microelectronics, TU Wien, A-1040 Wien, Austria
° IMEC, B-3001 Leuven, Belgium
† Infineon Technologies, Corporate Reliability Methodology, München, Germany

Abstract

Measuring the degradation of modern devices subjected to bias temperature stress has turned out to be a formidable challenge. Interestingly, measurement techniques such as fast-V_{th}, on-the-fly $I_{\mathrm{D,lin}}$, and charge-pumping give quite different results. This has often been explained by the inherent recovery in non-on-the-fly techniques. Still, all these techniques deliver important information on the degradation and recovery behavior and a rigorous understanding linking these results is still missing. Based on our detailed studies of the recovery, we propose a new measurement technique which allows the simultaneous extraction of two distinctly different components, a fast universally recovering component and a slow, nearly permanent component.

Introduction

The degradation in transistor parameters accumulated during bias temperature stress is commonly described by a shift of the drain current or threshold voltage as a function of stress time [1]. Once the stressing conditions are removed, rapid recovery sets in, making the measurement of the true degradation extremely challenging. Several measurement techniques have been developed and improved recently in order to determine the real shifts as accurately as possible, including ultra-fast V_{th} measurements [2], ultra-short pulse $I_{\mathrm{D}}V_{\mathrm{G}}$ techniques [3], on-the-fly (OTF) techniques [4]. From these investigations interface states and trapped positive charges have often been given as the most likely candidates causing degradation [2, 3, 5]. Consequently, additional techniques such as charge-pumping (CP) and DCIV techniques methods have been used to directly assess the interface state density [5, 6]. However, it is unclear how to relate the data obtained from the various techniques, whether they see two different or a single component in different forms, and – provided there is more than one – which component they actually measure [5, 7].

Analysis of Recovery

We propose a measurement technique based on the measure/stress/measure approach (MSM) *which simultaneously observes two distinctly different components* contributing to BTI: a large and recoverable component R which is extracted on top of a permanent component P. The existence of two components is consistent with a growing number of recent publications [5, 7, 8] and we link the results obtained from our methodology to reports in the literature.

Our technique is based on the *universality of BTI recovery* which has previously been observed for the recovery of the NBTI degradation in pMOSFETs [9, 10]. Here we show that this universality equally holds for NBTI/PBTI stress in pMOS and nMOS transistors: When relaxation data collected after different stress times $t_{\mathrm{s},i}$ are normalized to the last respective stress value and the relaxation times are normalized to the last stress times, all normalized data line up on a single universal curve, described by the universal relaxation function eq. (2). In conventional MSM measurements the last stress values are not directly available due to the immediate recovery but can be extrapolated using this universality [10]. Recently acquired more detailed data strongly indicate [11] that in addition to this universally recovering component R an additional permanent component P exists, which was only vaguely detectable in the previously available relaxation measurements [10]. However, due to limited data, the P component was incorrectly assumed to follow a power-law time dependence in [10].

In order to study this significant component more thoroughly, we have collected a large number of detailed relaxation data under various bias, duty factor, and temperature conditions using devices with plasma nitrided oxides from two foundries (EOT = 1.4 nm and 2.2 nm). Furthermore, two different measurement techniques (on-the-fly $I_{\mathrm{D}}@V_{\mathrm{G}} \approx V_{\mathrm{th}}$ [12] and fast-direct V_{th} [2]) were employed, yielding basically the same result.

Our measurement sequence is illustrated in Fig. 1: N relaxation sequences are collected in a single measurement (at given V_{G} and T) on a single device by interrupting the stress phase N times, with a last long relaxation sequence in the end [2, 10]. In an optimization loop the two components R and P are extracted by first subtracting the yet to be determined permanent parts P_i of all N individual relaxation sequences following eq. (1) and mapping the remainder of the relaxation sequences on the universal relaxation curve eq. (2). This requires the extraction of $N+2$ model parameters, 2 for the universal component R and N for the detailed information on P. These N samples of the permanent component P_i at each stress time $t_{\mathrm{s},i}$ allow a *detailed study of the time dependence of P*.

Discussion

The application of this technique is shown in Fig. 2 using $N = 9$ relaxation sequences obtained by fast direct ΔV_{th} measurements [2] ($t_{\mathrm{M}} \approx 1\,\mu\mathrm{s}$). The relaxation data clearly levels off, indicating the existence of a permanent or slowly relaxing component. Although the extracted stress-time dependence of P may *appear to initially follow a power-law*, it *clearly shows signs of saturation* at longer stress times, which is *fundamental for lifetime extrapolation*. We remark that if P relaxes slowly rather than being permanent, the extracted values are suitable averages $P_i = \langle P(t_{\mathrm{s},i}, t_{\mathrm{r}})\rangle$. This may have a consequence on the observed saturation. Note that no explicit functional forms of

1-4244-0439-X/07/$25.00 © 2007 IEEE

$R(t_s)$ and $P(t_s)$ are assumed during the parameter extraction, the only explicit expression we make use of is the universal relaxation function eq. (2) which excellently fits all the data discussed here. Interestingly, the same methodology can be applied to PBTI stress of pMOS and nMOS devices as well (Fig. 3). Although the extracted components are considerably smaller compared to the pMOS/NBTI case, the similarities are striking. Note that both stress conditions result in a *negative shift of the threshold voltage*.

The robustness of the extracted components is shown in Fig. 4 where from the measured $N = 9$ relaxation sequences only the first $M \leq N$ are considered, resulting in virtually no change in the extracted components. The same applies if from the fast ΔV_{th} data the first few decades ($1\,\mu s - 1\,ms$) are dropped, thereby emulating a slower measurement equipment. Provided $t_M < 10\,ms$, virtually the same results are again obtained. *We therefore conclude that the proposed extraction procedure is robust*.

In order to demonstrate that the extracted *component P has a physical meaning*, we accelerate recovery by applying a positive bias during relaxation [7]. As shown in Fig. 5, moderate positive bias accelerates recovery astonishingly close to the value of the extracted *P*, seemingly independent of the applied bias during relaxation. When a larger positive bias is applied, however, the degradation *increases* again, a peculiar fact observed in both technologies and unexplained by available NBTI models.

Also interesting is the observation that for a proper choice of stress voltages the *extracted P component is similar* for the two completely different devices and setups, *with R significantly larger for thinner oxides*, cf. Fig. 6. Another experiment demonstrating that the recovery after bias temperature stress is dominated by the recovery of *R* is given in Fig. 7, where the same device is subjected to the same stress/relax experiment twice, with only a short recovery in between. In the second run no change in *P* is detectable while the extracted *R* is similar to the one extracted from the first run.

A crucial aspect of BTI degradation which also has eluded satisfactory explanation so far is its duty-factor (DF) dependence. It has been previously shown [13] that the ratio of DF/DC degradation is a strongly non-linear function of the DF. Replacing the stress phases in our methodology by AC stress phases, the DF dependence of *P* and *R* can be extracted revealing that *the strong DF dependence is due to fast recovery of R* with *P* being either permanent or only slowly relaxing (Fig. 8). Furthermore, the functional form of the DF dependence can be estimated from the universal relaxation law eq. (2), assuming $t_{s,eff} = t_s \times DF/100$ and $t_r = t_s - t_{s,eff}$.

Another important aspect for lifetime projection is the temperature and voltage acceleration of the two individual components, which has been studied in detail with the last long recovery phase displayed in Fig. 9. For the temperature and voltage ranges investigated, no anomalies have been observed, resulting in excellent fits to the model eq. (1). The extracted activation energies of *R*, *P*, *B*, and β are shown in Fig. 10. The activation energy of *R* extracted after different stress times follows the same Arrhenius law while the activation energy of *P* depends on the stress time, making *P essentially non-Arrhenius* [7]. The parameters of the universal relaxation function eq. (2) are Arrhenius (B) and constant (β). Interestingly, a temperature independent β contrasts with predictions of dispersive hydrogen transport models [10, 12]. As shown in Fig. 11, the temperature and voltage dependence of *P* can be well described by the analytic expression derived in [7]. This expression has been used to account for the time, voltage, and temperature dependence of interface states as observed in CP measurements.

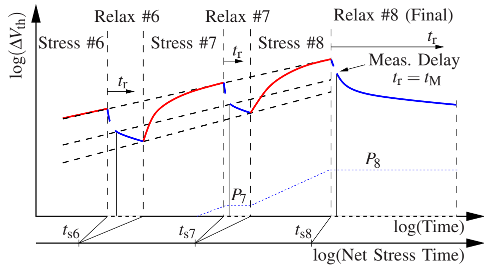
With the extracted model parameters for eq. (1) it is possible to study the influence of the measurement delay on the observed degradation in detail. Fig. 12 shows that the data is bound between $S = R + P$ (the extrapolated 'true' degradation) and *P*, and depends in a complex manner on the delay time, temperature, and bias conditions. This figure also nicely explains why so many different NBTI exponents have been reported in literature. In particular, it is found that the commonly used power-law is an approximate concept, valid only over a few decades in time, and should therefore be used with care for lifetime extrapolation.

Finally, our method is compared with a method suggested in [6] for HfO_2 layers. There it was suggested that the slow component (our *P*) is hidden by a large fast component (our *R*). By subtracting an early value of the total threshold voltage shift $S_M(t_0, t_M)$, the slow component may be retrieved under the assumption that *R* is nearly saturated at $t_s = t_0$ and $P(t_0) \ll R(t_0)$. Thus a reasonable approximation of *P* may only be obtained with a good choice of t_0 and t_M, the determination of which, however, is not straightforward.

Conclusions

We show that by recording several sequences of V_{th} relaxation at different stress times in a single measurement, the full stress and relaxation behavior can be reconstructed and separated from the permanent V_{th} degradation. The method is based on the universality of recovery and is shown to be valid for negative and positive temperature stresses. We demonstrate how the extracted relaxation model can be used to accurately study the influence of the measurement delay, temperature and voltage acceleration. In particular, our results suggest that standard MSM or OTF measurements plotting ΔV_{th} vs. t_s are of limited value, as they merely show the *combination of both R and P components*, each with their own different acceleration behavior.

[1] D. Schroder, MR **47**, 841 (2007).
[2] H. Reisinger *et al.*, in *IRPS* (2006), pp. 448–453.
[3] C. Shen *et al.*, in *IEDM* (2006), pp. 333–336.
[4] M. Denais *et al.*, in *IEDM* (2004), pp. 109–112.
[5] S. Mahapatra *et al.*, in *IRPS* (2007), pp. 1–9.
[6] A. Neugroschel *et al.*, in *IEDM* (2006), pp. 1–4.
[7] V. Huard *et al.*, MR **46**, 1 (2006).
[8] A. Haggag *et al.*, in *IRPS* (2007), pp. 452–456.
[9] M. Denais *et al.*, in *IRPS* (2006), pp. 735–736.
[10] T. Grasser *et al.*, in *IRPS* (2007), pp. 268–280.
[11] T. Grasser *et al.*, in *ESSDERC* (2007), pp. 127–130.
[12] B. Kaczer *et al.*, in *IRPS* (2005), pp. 381–387.
[13] R. Fernandez *et al.*, in *IEDM* (2006), pp. 1–4.



Relaxation model:

$$S_M(t_{s,i}, t_r) = \frac{R(t_{s,i}, t_M)}{r(t_M/t_{s,i})} r(t_r/t_{s,i}) + P_i \quad (1)$$

Universal relaxation law:

$$r(t_r/t_s) = r(\xi) = (1 + B\xi^{\beta})^{-1} \quad (2)$$

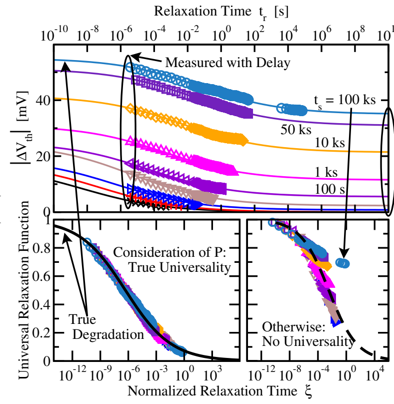
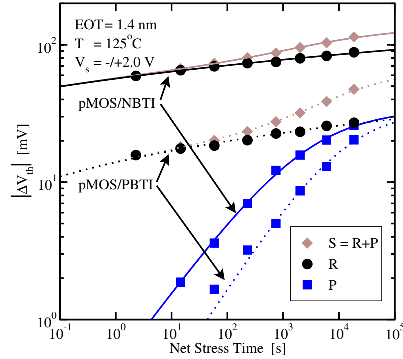
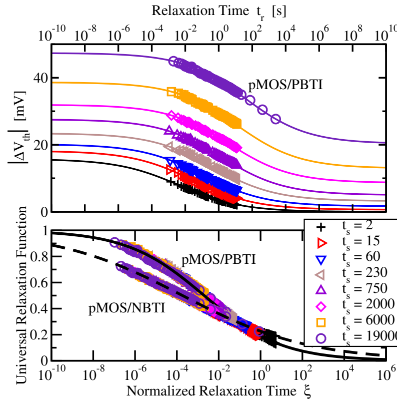


Fig. 1: Schematic view of the employed stress/relaxation sequences [12]. The stress is interrupted N times to record $N-1$ short and 1 long final relaxation sequence on the relative time scale $t_r = t - t_s$. Indicated is the permanent/slowly relaxing component P. The $N+2$ parameters B, β, and P_i ($i = 1 \ldots N$) are determined by matching eq. (1) to all N relaxation sequences.

Fig. 2: Extraction of the $N+2$ parameters of eq. (1) using fast threshold voltage measurements with $t_M \approx 1\,\mu$s. **Left**: The extracted model (lines) is in excellent agreement with the measured data (symbols). The universally relaxing component R on an absolute time scale on top of the permanent component P (top). The universality of R is clearly visible (bottom-left). Without considering P, large stress times break the universality (bottom-right). **Right**: The original measured data $S_M(t_s, t_M \approx 1\,\mu s)$, the extracted recoverable and permanent components R and P. The total degradation S is obtained as the sum of R and P by extrapolation of $R(t_s) = R(t_s, t_M = 0)$. Also indicated is the relaxation data on a relative time scale $t_{s,i} + t_r$. The recoverable component R can be well fit by a power-law and $n\log(1 + At_s)$ (used in the following) while P closely follows $P_{max}/(1 + (t_s/\tau)^{-\alpha})$ (see [7]).





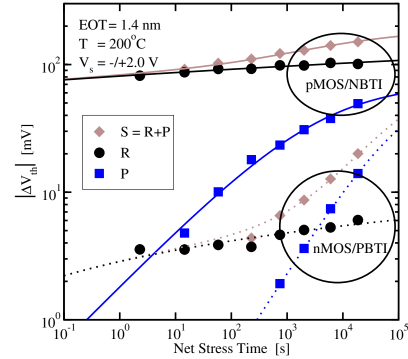
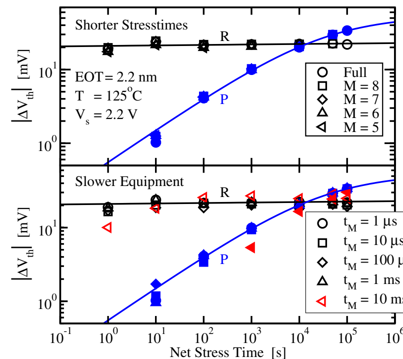
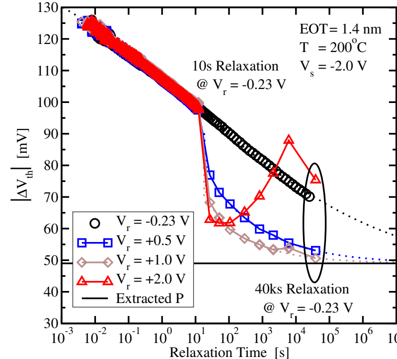


Fig. 3: A universally relaxing component is also present during PBTI stress in pMOS transistors (left) as well as in nMOS transistors (right) and the presented methodology can be applied to these cases as well. (symbols: data, lines: model). These samples have a much smaller EOT compared to the devices of Fig. 2, a considerably larger R but a comparable P component. **Left**: Comparison of NBTI and PBTI stress in a pMOS. The recoverable component during PBTI stress is considerably smaller but of similar shape while the permanent component is similar. **Middle**: The detailed relaxation measurements for the extraction of the left figure. An excellent fit of the model is obtained with relaxation after PBTI stress being somewhat faster than after NBTI stress. **Right**: Comparison of pMOS/NBTI and nMOS/PBTI stress. In the nMOS the recoverable component is very small and the degradation dominated by P from an early stage. Again, the permanent component behaves similarly.





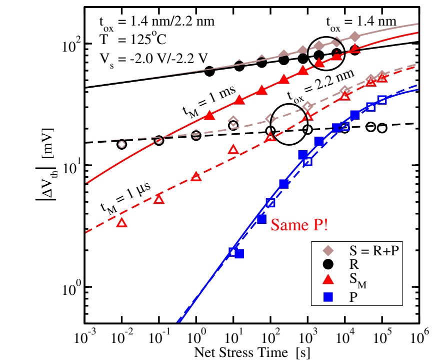


Fig. 4: Top: Inclusion of only the first M of the $N = 9$ relaxation sequences into the extraction procedure gives nearly indistinguishable results. **Bottom**: By dropping the first few decades of the $t_M = 1\,\mu$s measurement, a slower measurement equipment is emulated. For an initial measurement delay of $t_M < 10$ms again nearly indistinguishable results are obtained, proving the robustness of the algorithm.

Fig. 5: Application of positive bias voltages during the last long relaxation to accelerate recovery. Moderate positive voltages $V_G \leq +1$V enhance the recovery down to a bias-independent value extremely close to the extracted permanent component P (49mV@$t_s = 18$ks, cf. Fig. 3 right), indicating that P is a real, physically based quantity. At larger positive biases, degradation is observed during recovery.

Fig. 6: Comparison of the extracted components for two completely different technologies with different EOT, different measurement setups, and different stress/relaxation patterns (cf. Fig. 2 and Fig. 3). The thinner oxide has a considerably larger R but the same P compared to the thicker oxide. Due to the larger R, the overall degradation $S = R + P$ is dominated by R, while the thicker oxides display a transition from an R to P dominated regime.

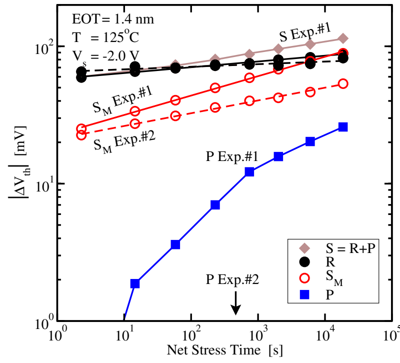
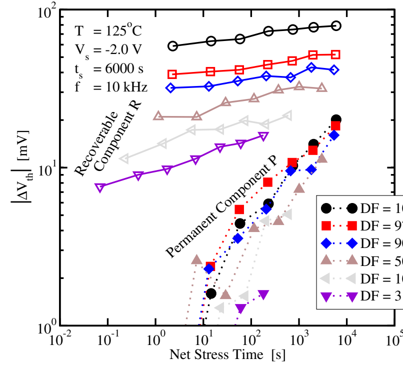


Fig. 7: The same device subjected to the same stress/relax experiment twice (solid vs. dashed lines). The initial V_{th} of Exp.#2 already contains the last P value of the Exp.#1 but no detectable change in P is observed during Exp.#2. While the as-measured values are clearly different due to the change of P in Exp.#1, the extracted R components are very similar.



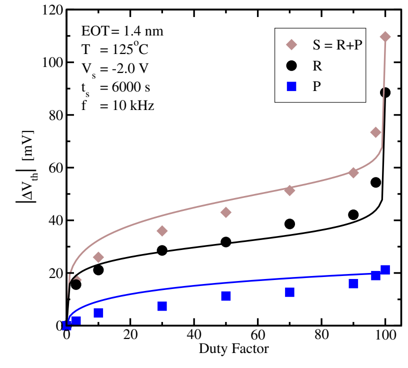


Fig. 8: Duty factor (DF) dependent unipolar AC stress experiments to determine the respective contributions of R and P. **Left**: While R(DF) shows a pronounced DF dependence, P(DF) is roughly similar, indicating that P is constant or possibly very slowly relaxing. **Right**: The values of S, P, and R after $t_s = 6000$ s stress as a function of the DF. A reasonable fit to the measurement data is obtained assuming $P_{AC} \approx P_{DC}(t_{s,eff})$, $R_{AC} \approx R_{DC}(t_{s,eff})r(\xi_{eff})$ with $t_{s,eff} = \alpha t_s$, $\xi_{eff} = (1-\alpha)t_s/(\alpha t_s)$, $\alpha = DF/100$, and $B_{AC} \approx B_{DC}/2$. The deviation in P_{AC} indicates that P might be slowly relaxing.

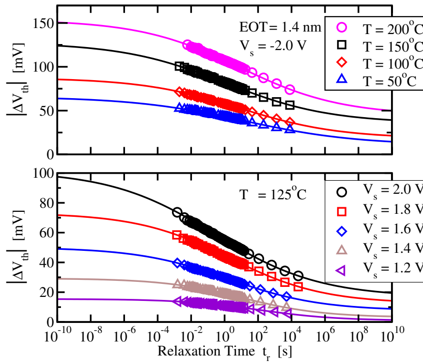


Fig. 9: Top: Temperature dependence of the last measured relaxation sequence after $t_s = 20000$ s. **Bottom**: Same for the stress voltage dependence. An excellent fit for all bias and temperature conditions is obtained using our procedure.

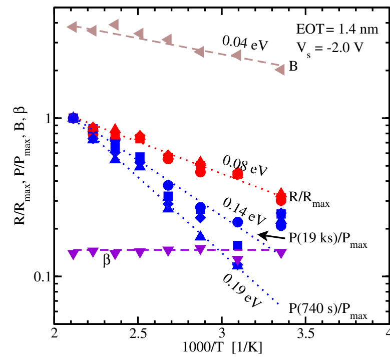


Fig. 10: Arrhenius plot showing the (apparent) activation energies for the P and R component obtained at different stress times ($t_s = 740 \dots 19000$ s) in addition to B and β required for the universal relaxation function eq. (2). While β is constant, R and B are Arrhenius but P is not.

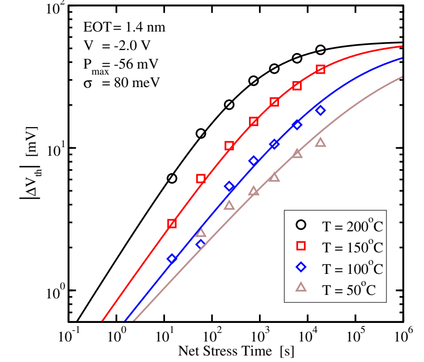
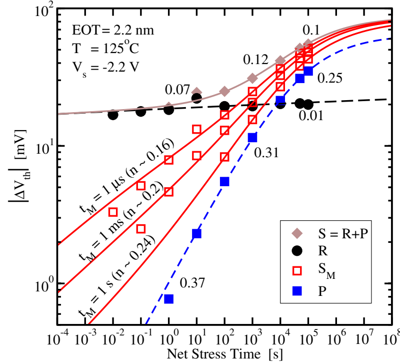


Fig. 11: Fit of the permanent component with the expression given in [7] for CP data, $P = P_{max}/(1+(t_s/\tau)^{-\alpha})$. With $\alpha = \sigma/k_B T$, $\tau = \tau_0 \exp(E_d/k_B T)$ and $E_d = E_{d0} + k(V_s - V_{th}(T))$ an excellent fit is obtained with very similar parameter values as in [7], suggesting a possible correlation between P and the interface state density D_{it}.



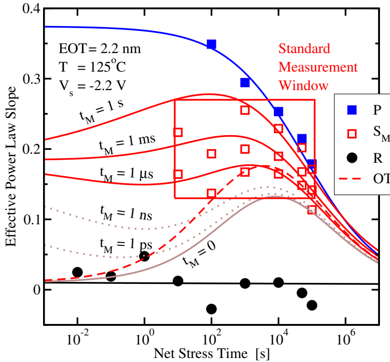


Fig. 12: Influence of the measurement delay on the observed NBTI behavior (open symbols: data, closed symbols: extracted R and P, lines: model). **Left**: The measurement result is bound by S and P. Depending on t_M a broad range of 'effective' power-law slopes may be observed, in agreement with [3] (limiting values given next to the model lines). **Right**: The effective power-law slope as a function of the measurement delay, defined as $d\log(S_M)/d\log(t_s)$. Irrespective of the fact that within a few decades, e.g. $t_s = 1 \dots 10^4$ s, the data can be apparently fit by a power-law, a power-law does not seem to be the best choice to represent the time behavior of NBTI, which in a complicated way depends on the interplay between R and P and thus on the measurement delay and the temperature (ratio R vs. P). For the OTF line it was assumed that the occupancy of interface states P is twice as high as during MSM measurements.

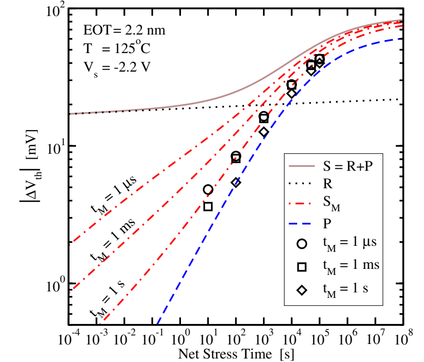


Fig. 13: For HfO_2 layers it was suggested [6] to subtract the initial 'hole trapping' component from the measurement data to extract the 'standard NBTI' as $P(t_s) \approx S_M(t_s, t_M) - S_M(t_s = t_0, t_M)$ with $t_0 = 1$ s. The accuracy of this approach requires $dR/dt_s \approx 0$ and $P(t_0) \ll R(t_0)$ which is also roughly fulfilled for nitrided oxides and reasonable approximations of P may be obtained by adjusting t_M and t_0 (symbols, lines: model from Fig. 12 (left)).