Gate-Sided Hydrogen Release as the Origin of "Permanent" NBTI Degradation: From Single Defects to Lifetimes

T. Grasser,[◇] M. Waltl,[◇] Y. Wimmer,[◇] W. Goes,[◇] R. Kosik,[◇] G. Rzepa,[◇]
H. Reisinger,[•] G. Pobegen,[‡] A. El-Sayed,[†,◇] A. Shluger,[†] and B. Kaczer[○]

[◇] *TU Wien, Vienna, Austria* [•] *Infineon Munich, Germany* [‡] *KAI, Villach, Austria* [†] *UCL, London, UK* [○] *IMEC, Leuven, Belgium*

Abstract

The negative bias temperature instability (NBTI) in pMOS transistors is typically assumed to consist of a recoverable (R) and a so-called permanent (P) component. While R has been studied in great detail, the investigation of P is much more difficult due to the large time constants involved and the fact that P is almost always obscured by R. As such, it is not really clear how to measure P and whether it will in the end dominate device lifetime. We address these questions *by introducing a pragmatic definition of P*, which allows us to collect long-term data on both large and nanoscale devices. Our results suggest that *(i)* P is considerably smaller than R, *(ii)* that P is dominated by oxide rather than interface traps and therefore *(iii)* shows a very similar bias dependence as R, and finally *(iv)* that P is unlikely to dominate device lifetime. We argue that a *hydrogen-release mechanism from the gate-side of the oxide*, which has been suspected to cause reliability problems for a long time [1–6], is consistent with our data. Based on these results as well as our density-functional-theory (DFT) calculations we suggest a microscopic model to project the results to operating conditions.

Introduction

The experimental characterization of NBTI has always been a challenging problem. While during the last decade a lot of effort has been put into developing ultra-fast characterization methods to assess mostly the recoverable component R, characterization of P faces the opposite problem of very large time constants which are difficult to access under normal experimental conditions. It has however been suggested that just like R, P is also thermally activated as it can be removed by a bake at 350°C [7, 8]. Various attempts have been made at measuring P, for instance using charge pumping [9, 10], DCIV [11], or spin-dependent recombination [12, 13], all of which should be very sensitive to interface traps (P_b centers), the most commonly suggested culprit. These measurements have shown that such interface traps do certainly contribute to P [10, 14] and that they are very likely silicon dangling bonds at the interface (P_b-centers) [12, 13]. It has been suggested, however, that in addition to interface states, hydrogen-related donor-like traps are created [1, 2] which could dominate P at longer times [3]. As the conversion of these data into ΔV_{th} is difficult, the magnitude and composition of P relative to R has remained controversial. As such, not too much is known about the temperature- and bias-dependence of P, which makes currently available lifetime extrapolation methods questionable.

Experimental Method

As summarized above, P is difficult to measure directly due to traps with very large relaxation times contribution to R, which overshadow P. P is thus typically assessed using a different method, the results of which must then be converted to ΔV_{th}. In order to measure P and R *simultaneously*, we employ 10 I_D/V_G sweeps into accumulation after a regular 100 s recovery [15], since it has been repeatedly shown that such bias sweeps remove a significant fraction of traps contributing to R [10, 16–19]. The remaining ΔV_{th} is henceforth *pragmatically called P*. The biasing scheme of Fig. 1 is first applied to large devices (10 μm $\times$ 10 μm) to study the average response of a large number of defects, as well as to nanoscale devices (150 nm $\times$ 100 nm) to study the creation and the annealing of individual defects. To access a wide temperature range (125-350°C), we use packaged devices of a 130nm commercial technology [20] (2.2nm SiON) in multiple computer-controlled furnaces.

Results Large Devices

While the P obtained using this pragmatic method almost certainly contains a contribution from R [19], P amounts to only 10% of the total ΔV_{th} measured at 1 μs even after a stress of 200 ks, see Fig. 2. So if the *real P* is significantly different, it is very unlikely to make a sizable contribution to NBTI. This P shows the expected dependence on the readout voltage (Fig. 2), which is typically associated with the changing occupancy of interface states. Utilizing the I_D/V_G sweeps, we can extract P as a function of the readout V_G, see Fig. 3. Typically, the first up-sweep from inversion to accumulation (typically from -0.6 V to $+1$ V) has the strongest impact. Also, P is smaller during the down-sweeps and we will take $P_{min} = \min(P)$ as a measure for P. It can be seen that the $V_G = V_{read}$ dependence of P does not show a peak inside the scanned region, already indicating that P_b centers are not the only contributor to P. Furthermore, P is independent of the duty factor (Fig. 4). Also shown is the impact of a bake cycle up to 350°C (see inset) which recovers most of the degradation and results in similar degradation during re-stress [7, 8]. Fig. 5 shows that the I_D/V_G sweeps have a strong impact mostly in depletion/accumulation (electron injection).

Next, the impact of the bake step is investigated in more detail. As can be seen in Fig. 6, while the bake seems to restore the ΔV_{th} dynamics, it does not fully restore V_{th}. In particular, baking introduces an additional contribution to P. This contribution seems to depend mostly on the bake itself. The peculiar finding that also a bake at 350°C/0 V causes degradation is studied more closely in Fig. 7 using various voltages during bake (temperature profile from Fig. 4). It can be seen that P can be cycled between different levels which depend on the voltage during bake. We stipulate that this experiment reveals P in an accelerated form rather than being a new effect. These results also clearly demonstrate that the *maximum obtainable P depends on the gate bias during operation*, similarly to R, where this effect is due to the "active area" in the band diagram accessible during stress [21].

Results Small Devices

In small devices, the impact of individual defects is visible in the I_D/V_G sweeps [23]. In addition to the simplest case of a fixed positive charge, which introduces a bias-dependent offset into the $P(V_G)$ curves, two types of defects are of interest: first, fast interface states will contribute to P until the Fermi-level $E_F(V_G)$

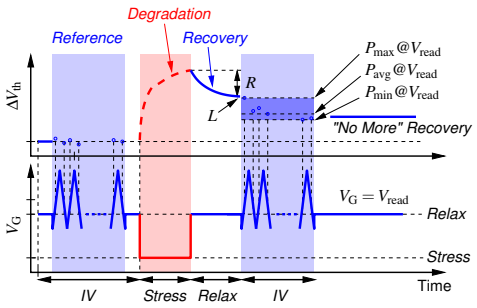

Fig. 1: Measurement sequence to determine the "permanent" component P. Prior to stress, 10 up/down I_D/V_G sweeps are recorded as a reference. After a stress of typically $t_s = 10\,s$ to $1\,ks$ and a recovery phase of typically $t_r = 100\,s$, another set of 10 I_D/V_G sweeps is recorded. Each up/down sweep is used to determine a "permanent" threshold voltage shift P. A significant amount of charge is removed during these sweeps, resulting in basically flat recovery.

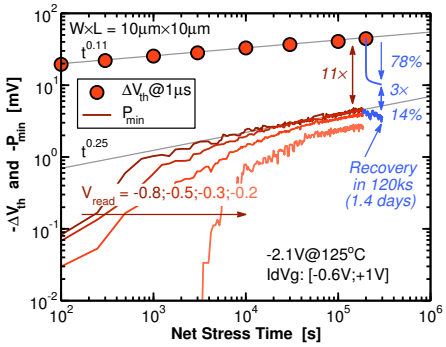

Fig. 2: Comparison of the standard ΔV_{th} shift measured with a delay of $1\,\mu s$ against P_{min}, which is about a factor of 10 smaller. As expected, P_{min} depends on the read-out voltage, which is typically explained by the presence of interface states which have a bias-dependent occupancy. P_{min} is nearly permanent and shows only little recovery (blue lines) compared to the regular ΔV_{th}. A power-law approximation for ΔV_{th} and P_{min} is also shown (gray lines).

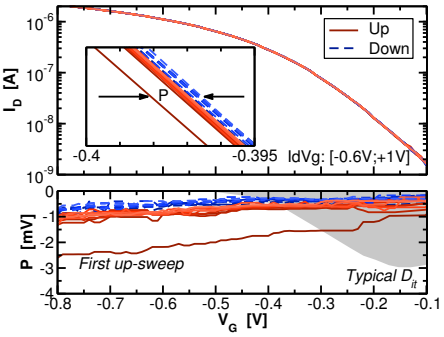

Fig. 3: From the 10 up and down sweeps (**Top**), P is extracted as a function of V_G (**Bottom**). Typically, P is significantly larger in the first up-sweep during which a significant portion of the trapped charges is annihilated/recovered. All subsequent sweeps only show spontaneous capture and emission. Also shown is a typical location of D_{it} [22] where a strong impact of interface traps would be expected (0.25 eV above the valence band edge).

reaches the trap-level E_T, see Fig. 8. On the other hand, oxide traps can be much slower, resulting in a hysteresis between the up and down sweeps. Furthermore, stochastic transitions between the levels would be expected (Fig. 9).

The evolution of P during a 77 day study is shown in Fig. 10. The 10 s/100 s stress/relax cycles were interrupted three times for a long recovery phase (1-3 days, $-0.3\,V@125°C$) where *no recovery was observed* and 11 times for a bake. Most changes in P are simply RTN, probably due to oxide traps with very large time constants (days). Also shown is the evolution of R, which clearly shows volatile oxide traps [24]. Similar results are obtained on a different device for 1 ks/100 s stress/relax cycles (Fig. 11). During these long-term time-dependent defect spectroscopy (TDDS) studies the noise in the recovery traces tended to increase. Interestingly, a bake at $+1\,V@350°C$ was found not only to nearly restore the initial V_{th} but also the initial noise level, see Fig. 12.

The "permanent" changes in P are analyzed in Figs. 13 and 14. Various types of behavior were found, most of them consistent with *oxide traps/charges* rather than interface states. However, interface states would have a trap level distributed around 250 mV above E_V (see Fig. 3), which is at the border of our experimental window, so about 50% of interface states would look like fixed charges in our experiments. In any case, no clean interface trap signal could be detected, confirming previous results which suggest that P is dominated by *oxide defects* [3], likely trapped hydrogen [2].

Modeling and Extrapolation

Our results are clearly consistent with the previously suggested hydrogen-release mechanism following hot electron injection [3, 25]. During NBTI, however, a slightly different mechanism for H release needs to be found. DFT calculations show [26–30] that in amorphous SiO_2 protons can be trapped in various configurations. A schematic model based on this result is shown in Fig. 15: First, protons can bind to bridging oxygens [28] with a wide distribution of energy levels (1 eV). Depending on the gate voltage during stress, the E_T of some traps are moved above E_F, are neutralized and released as H^0 and *quickly migrate from the gate to the channel* occupying trapping sites that were previously not available at $V_G = V_{th}$, thereby forming P. Some of those H^0 may also depassivate P_b centers, thereby also giving a convenient explanation for the puzzling observation that during NBTI stress such strong Si-H bonds ($E_B = 2.5\,eV$ for the direct removal of H [31, 32]) are broken [33]. Secondly, when the local strain in the structure is larger, H^0 can remain attached to their bridging oxygens even after having been neutralized [29, 30]. They can only be released over an 1-1.5eV large barrier and are thus immune to neutralizing I_D/V_G sweeps. Our data suggest that it is those trapped hydrogens which form the major contribution to P in addition to a smaller contribution of P_b centers [3, 34–36]. To explain the additional increase in P at 350°C, we assume that additional H can be released from the poly-gate over a barrier of 2.5 eV (e.g. the Si-H binding energy [31, 32]), which then moves into the recently vacated H-trapping sites. Given the high diffusivity of hydrogen [37], the response of the system is purely *reaction*-limited [38].

The complete model is evaluated with good accuracy against the data in Fig. 16, showing first how P saturates at each stress temperature and then how the number of available defects can be increased at temperatures higher than 350°C. The re-distribution of hydrogen *towards the channel* is shown Fig. 17, consistent with nuclear reaction analysis (NRA) results [6]. Finally, an extrapolation towards 125°C demonstrates that P will unlikely dominate the degradation after 10 years, see Fig. 18. Note that the extrapolation was only done with respect to the temperature, as experiments were conducted mostly at $V_G = V_{DD} = -1.5\,V$.

Conclusions

Using a pragmatic definition of the "permanent" component of NBTI, we have studied P at the multi- and single-defect level using a long-time dataset. We found that just like the recoverable component R, P depends on a bias-dependent active region in the oxide. Also, even after 10 ks of recovery, a significant fraction of ΔV_{th} can be removed by a sweep into accumulation, with the remainder (P) being probably dominated by trapped hydrogen rather than P_b centers. Finally, it was shown that while P can provide a sizable contribution to ΔV_{th}, it does *not appear to dominate the lifetime* at use-conditions.

The research leading to these results has received funding from the Austrian Science Fund (FWF) project n°26382-N30, the European Community's FP7 project n°619234 (MoRV), and the Intel Sponsored Research Project n°2013111914. The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC) and the UK's national high-performance computing service HECToR and Archer via the Materials Chemistry Consortium (EPSRC EP/F067496). Valuable discussions with V. Afanas'ev, A. Stesmans, J. Stathis, E. Wu, and E. Cartier are gratefully acknowledged.

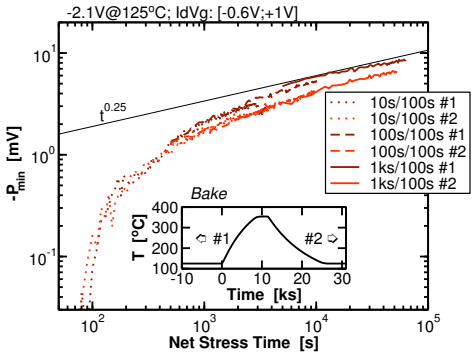

Fig. 4: The measured P is independent of the duty cycle, shown for 3 devices stressed with 3 different stress/recovery times (t_s/t_r=10/100, 100/100, and 1000/100), which nicely overlap. *This is never the case for the full threshold voltage shift* ΔV_{th}. After experiment #1, the devices were baked at 350°C (inset), which lead to a slightly smaller P in run #2. However, the three stress/recovery patterns were consistent.

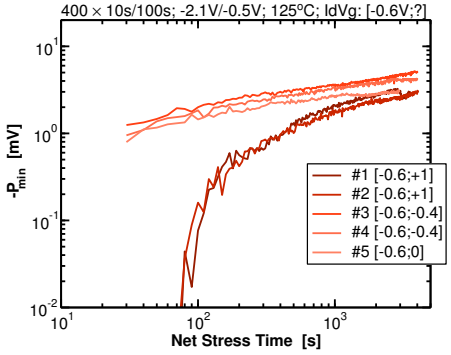

Fig. 5: The range used for the I_D/V_G sweeps has a strong impact on the measured "P". For a wide sweep range [-0.6;+1] (electron injection), a large amount of charge is removed and a small P is obtained. For a narrower sweep range [-0.6;-0.4] (around V_{th}), only little charge is removed, giving a large "P". The experiments were performed subsequently with a bake step inbetween (0 V/350°C).

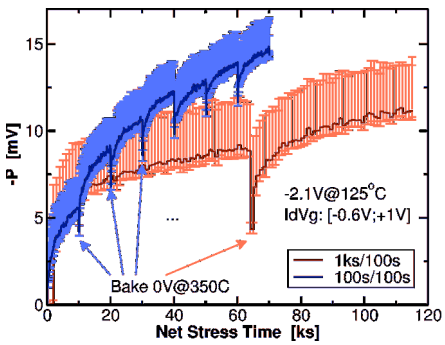

Fig. 6: While a 1h bake with $V_G = 0$ V at 350°C 'resets' most of the dynamics (cf. Fig. 4), it also leads to an additional offset/degradation. This seems to be a result of the bake step itself, meaning that contrary to previous assumptions, $V_G = 0$ V both recovers some fraction of the degradation but also leads to additional stress of the device at 350°C. Shown are $\langle P \rangle = P$ (solid dark lines) while the error bars give P_{min} and P_{max}.

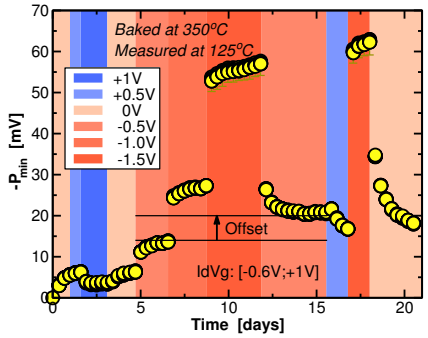

Fig. 7: Even baking at V_G=0 V causes as buildup of positive charge and a degradation of about 8 mV, while baking at more positive voltages removes some of that charge. Every datapoint above is recorded between the bake cycles. For every bake/stress voltage there appears to be a well defined maximum degradation (at least from a life-time perspective). Except for a small offset (one shown), most of this built-up charge can be easily cycled at 350°C. We speculate that even this offset could be removed after longer anneal times.

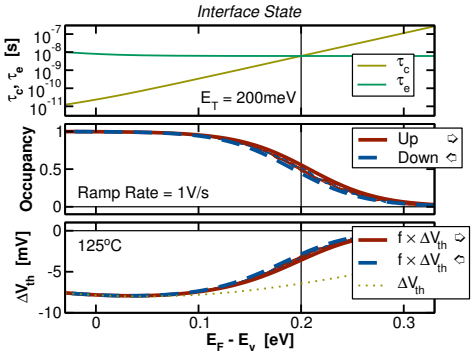

Fig. 8: Calculated response of a single interface state to I_D/V_G sweeps according to Shockley-Read-Hall statistics ($\sigma = 10^{-16}\,\mathrm{cm}^{-2}$). **Top**: The capture and emission times. **Middle**: The occupancy of the trap quickly follows the sweep and the measurement only records the averages and there is practically no hysteresis between the up and down sweep. The averages over many sweeps are shown in bold. **Bottom**: Only when the trap is occupied, the charge at the trap position impacts ΔV_{th}. Since the trap is very fast, the measurement only sees the average.

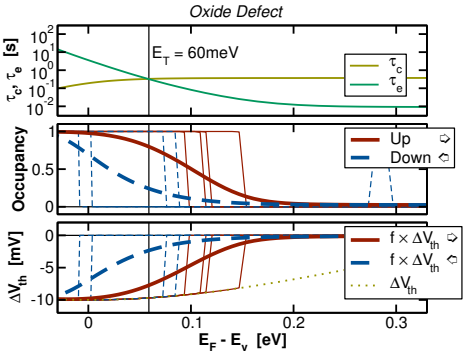

Fig. 9: Same as Fig. 8 but using a four-state non-radiative multiphonon (NMP) oxide trap model [39]. Oxide traps have much larger capture and emission times (**Top**). As a consequence, hole emission occurs at a higher Fermi-level E_F than capture, resulting in a hysteresis in the average behavior (fat lines) (**Middle**). These emission events would be clearly visible in the experimental data.

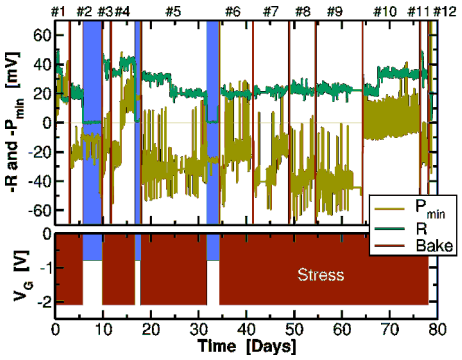

Fig. 10: A nanoscale device was stressed/recovered (10 s/100 s) for nearly 80 days at −2.1 V@125°C. Before each of the 12 blocks, the device was baked at 0 V/350°C. Excepting block #4 (which was after a bake at +1 V), only little build-up of P is observed during stress (red areas). During recovery (blue shaded area), P stayed constant for several days. Note the fluctuations in R ("volatility" [24]) as well as in P.

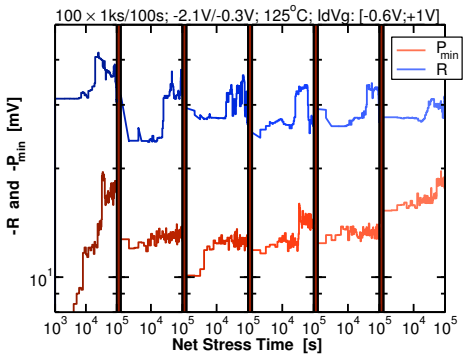

Fig. 11: Similar to Fig. 10, 6 runs with 100 repetitions of 1 ks stress followed by 100 s recovery and 10 I_D/V_G sweeps. The 6 runs were each separated by a bake at 0 V/350°C. As with Fig. 10, most changes in P are very slow random-telegraph-noise-like fluctuations while only occasionally abrupt large "permanent" changes are observed. Such "permanent" changes are analyzed in more detail in Fig. 13 and Fig. 14.

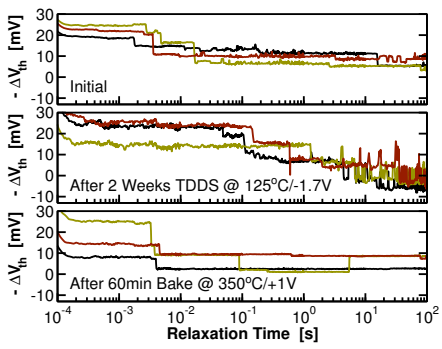

Fig. 12: A device was repeatedly stressed for 10 s at −1.7 V and recovered for 100 s at −0.8 V at 125°C for 2 weeks. Initially, the traces were relatively clean (**Top**). With continuing stress, the traces become increasingly noisy (**Middle**) due to the appearance of new defects. A bake at +1 V at 350°C was found to be able to much more efficiently remove these new defects (**Bottom**) than a bake at 0 V/350°C (not shown).

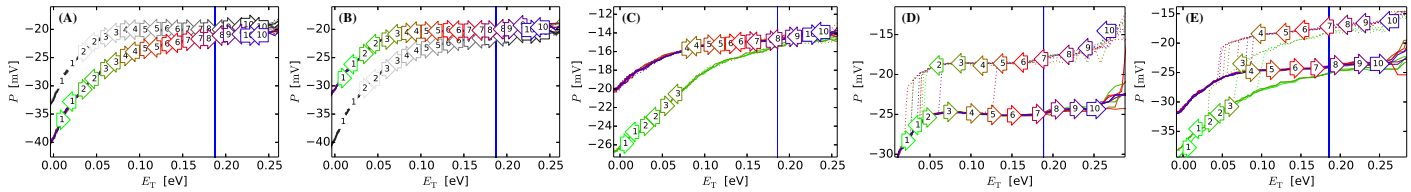

Fig. 13: I_D/V_G sweeps in a nanoscale device (here the device from Fig. 12) allow a clear identification of individual traps. The numbered arrows give the sweep direction and number while the gray version (wherever shown) corresponds to the previous stress/relax cycle. The vertical blue line corresponds to the read out voltage (-0.3 V). **(A)** This is the simplest case: during stress a fixed positive oxide charge is created. **(B)** The very same oxide charge is annealed at a later point. **(C)** Fixed oxide charge anneals (randomly) sometime during the sweeps in a bias-independent manner (here between sweep 3 and 4). **(D)** During stress a very slow switching trap is created with a trap level around 200 mV. **(E)** In general, multiple traps are active during the sweeps, here the traps of (C) and (D) plus another small one in the first up- and down-sweep.

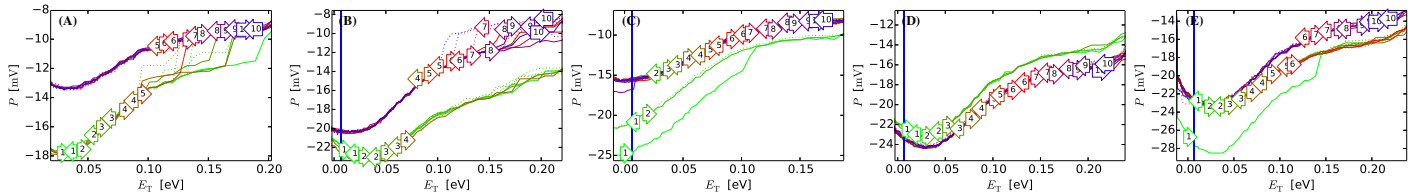

Fig. 14: Same as Fig. 13 but for a different device. **(A)** A switching trap repeatedly changes its occupancy until it completely anneals after up-sweep 5. **(B)** A fixed positive charge disappears in up-sweep 4. At the same time, a switching trap with $E_T \approx 0.15$ eV appears. This happened repeatedly and was not a coincidence. **(C)** A trap (likely the one from (A)) with $E_T \approx 0.1$ eV disappears in up-sweep 1, and another fixed positive trap in up-sweep 2. **(D)** During up-sweep 3 a positive charge is *captured*, most likely from the gate. **(E)** Switching trap (A) disappears in up-sweep 1 and another fixed positive trap in up-sweep 6.

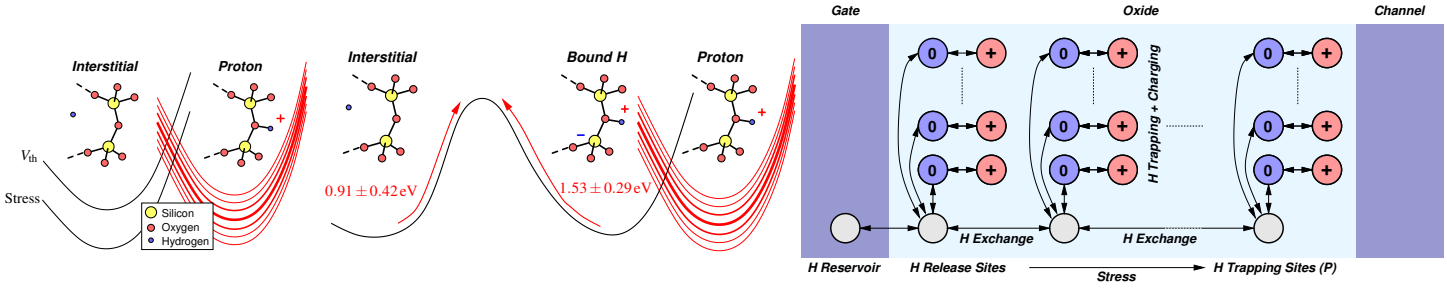

Fig. 15: Left: Depending on E_F and the configuration, interstitial H can capture a hole over an NMP barrier and attach to the bridging oxygen. When E_F is moved, the hole can be emitted and H becomes interstitial (shown is a case next to the gate). There is a broad distribution of energy levels ($\sigma \approx 1$ eV) and the NMP emission barrier can be made very small with suitably chosen gate bias. Hydrogens bound in such a configuration will be neutralized and released at the gate side during stress. Alternatively, such hydrogens will be released from the channel side for instance during I_D/V_G sweeps and a $+1$ V bake. **Middle**: When the Si-O bond is strained, H remains bound to the bridging oxygen even in the neutral configuration with a significant barrier to the interstitial position. H bound in such configurations will not be released during I_D/V_G sweeps provided the barriers are large enough, contributing to P (in red: DFT barriers [29]). **Right**: A schematic view of the H-release model. H can get trapped in two configurations (see **Middle**/**Right**) and the trapping sites are connected via barriers to the interstitial configurations. Redistribution of H in these configurations (diffusion) is very fast and not rate-limiting [37]. At $T \gtrsim 350$°C, additional H can be exchanged with the reservoir, separated by a 2.5 eV barrier (e.g. Si-H bonds [31, 32]).

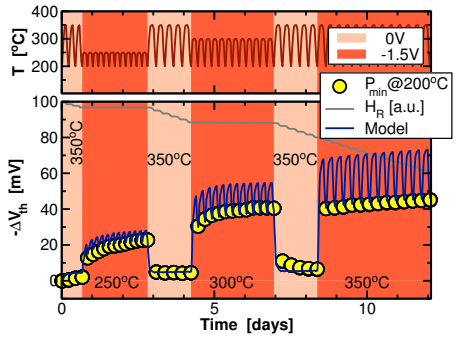

Fig. 16: The H-release model vs. data (**Bottom**) at various temperatures (**Top**: T profile). Each symbol is a measurement taken at 200°C between T ramps. At 350°C, additional H is supplied by the reservoir. This effect, however, is *irrelevant* under normal operation and only required to understand the 350°C data.

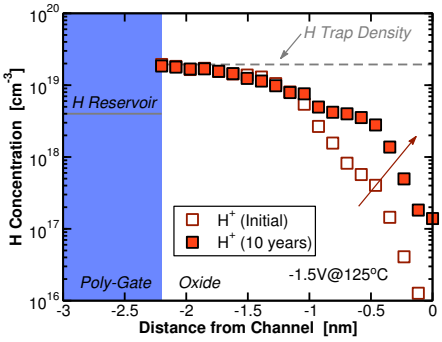

Fig. 17: Redistribution of H towards the channel during a 10 year stress with -1.5 V@125°C (cf. Fig. 18). No additional H is supplied from the reservoir close to these use conditions. For simplicity, a constant H trap density is assumed [6]. The noise is due to the stochastic algorithm employed in the model [40].

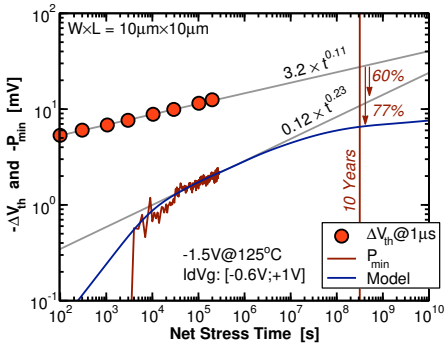

Fig. 18: Using only T acceleration, we find that P does *not* dominate the 10-year degradation value. Using a simple power-law extrapolation, P is about half of the total ΔV_{th}. Since P is limited by the amount of available H trapping sites, the H-release model predicts P to be only about 23%.

References

[1] E. Cartier *et al.*, ME **28**, 3 (1995). **[2]** J. de Nijs *et al.*, APL **65**, 2428 (1994). **[3]** E. Cartier, MR **38**, 201 (1998). **[4]** V. Afanas'ev *et al.*, PRL **80**, 5176 (1998). **[5]** M. Nelhiebel *et al.*, MR **45**, 1355 (2005). **[6]** Z. Liu *et al.*, ECS T. **35**, 55 (2011). **[7]** A. Katsetos, MR **48**, 1655 (2008). **[8]** G. Pobegen *et al.*, *IEDM* (2011), p. 27.3.1. **[9]** D. Ang *et al.*, EDL **26**, 906 (2005). **[10]** V. Huard *et al.*, MR **46**, 1 (2006). **[11]** A. Neugroschel *et al.*, *IEDM* (2006), p. 1. **[12]** J. Campbell *et al.*, T-DMR **6**, 117 (2006). **[13]** J. Ryan *et al.*, APL **96**, 223509 (2010). **[14]** D. Ang, EDL **27**, 412 (2006). **[15]** T. Aichinger *et al.*, JAP **107**, 024508 (2010). **[16]** J. Zhang *et al.*, T-ED **51**, 1267 (2004). **[17]** B. Kaczer *et al.*, *IRPS* (2008), p. 20. **[18]** D. Ang *et al.*, T-DMR **8**, 22 (2008). **[19]** T. Grasser *et al.*, *IRPS* (2011), p. 605. **[20]** H. Reisinger *et al.*, *IRPS* (2006), p. 448. **[21]** T. Grasser, MR **52**, 39 (2012). **[22]** L. Ragnarsson *et al.*, JAP **88**, 938 (2000). **[23]** J. Franco *et al.*, *IRPS* (2012), p. 5A.4.1. **[24]** T. Grasser *et al.*, *IRPS* (2015). **[25]** J. Zhang *et al.*, JAP **87**, 2967 (2000). **[26]** P. Blöchl, PRB **62**, 6158 (2000). **[27]** J. M. Soon *et al.*, APL **83**, 3063 (2003). **[28]** J. Godet *et al.*, PRL **99**, 126102 (2007). **[29]** A.-M. El-Sayed *et al.*, PRL **114**, 115503 (2015). **[30]** A.-M. El-Sayed *et al.*, PRB **92**, 014107 (2015). **[31]** A. Stesmans, PRB **61**, 8393 (2000). **[32]** L. Tsetseris *et al.*, T-DMR **7**, 502 (2007). **[33]** S. Mahapatra *et al.*, T-ED **60**, 901 (2013). **[34]** J. Stathis *et al.*, PRL **92**, 087601 (1 (2004). **[35]** Y. Roh *et al.*, J. Noncryst. Solids. **187**, 165 (1995). **[36]** E. Cartier *et al.*, APL **69**, 103 (1996). **[37]** D. Griscom, JAP **58**, 2524 (1985). **[38]** T. Grasser *et al.*, T-ED **61**, 3586 (2014). **[39]** T. Grasser *et al.*, *IRPS* (2010), p. 16. **[40]** D. Gillespie, J.Comp.Phys. **22**, 403 (1976).