Unambiguous Identification of the NBTI Recovery Mechanism using Ultra-Fast Temperature Changes

Thomas Aichinger
Kompetenzzentrum für Automobil- und Industrieelektronik (KAI)
Europastrasse 8, 9524 Villach, Austria
0043–4242- 34890-24, aichinger.thomas@gmx.at

Michael Nelhiebel
Infineon Technologies Austria
Siemensstrasse 2, 9500 Villach, Austria

Tibor Grasser
Christian Doppler Laboratory for TCAD
Institute for Microelectronics TU Wien
Gusshausstrasse 27-29, 1040 Wien, Austria

***Abstract*-We study temperature and bias dependence of V_{TH} and interface states recovery after NBTI (Negative Bias Temperature Instability) stress. By making use of in situ heated test structures, we are able to change the temperature quickly and reliably at any stage of the experiment while keeping the device bias conditions untouched. This tool enables us on the one hand (i) to bring identically processed devices to the same degradation level, by stressing them under the same bias and temperature conditions, and on the other hand (ii) to vary the temperature in a defined way during recovery. Additionally we also study the influence of gate bias switches at constant temperature. Out of those experiments we have discovered that recovery acceleration can be observed either by increasing the temperature or by switching the gate bias for a short period of time towards accumulation. Since Charge Pumping (CP) measurements before stress and after recovery indicate that interface states cannot be made responsible for this acceleration, we suggest the neutralization of NBTI induced positive oxide traps to be the dominant recovery mechanism. However, if this is the case, we have to consider inelastic phonon-assisted tunneling in order to explain the temperature acceleration at fixed gate bias.**

Keywords-NBTI recovery, polyheater, interface states, oxide traps, charge pumping

I. INTRODUCTION

During NBTI oxide trap and/or interface state creation is accelerated by high fields, free substrate holes and elevated temperatures [1-4]. In a PMOS transistor such traps are typically positively charged for $V_{GS}=V_{TH}$, leading to a negative threshold voltage shift after stress. However, until now the relative amount and correlation of different contributions to total threshold voltage shift and recovery is still a controversial point which requires an unambiguous clarification in order to formulate reliable degradation and recovery models. In the past many experimental and theoretical attempts have been made in order to explain for example the log-like time dependence of V_{TH} recovery over many decades of time or the bias and temperature dependent acceleration of degradation. From these studies two main NBTI mechanisms have emerged, namely charging and recharging of NBTI induced oxide defects by tunneling from carriers from the silicon substrate [2-3], and hydrogen diffusion, which has often been attributed to the passivation and depassivation of defects by diffusing hydrogen species within the oxide and at the interface [4]. From a theoretical point of view such fundamentally different physical phenomena should show completely different dependencies on experimental parameters like gate bias, temperature and recovery time. Previous experiments dealing with bias switches during recovery [2-3], demonstrated that hydrogen reaction and diffusion cannot be made responsible alone for degradation and recovery. As a solution carrier trapping via elastic tunneling was suggested. However, an elastic tunneling process would be definitely too fast to explain long term recovery over many decades in time (especially for ultra thin oxides). A recently published technique allowing fast temperature switches on single devices using in situ heated poly silicon structures enabled us for the first time to bring identically processed devices to the same degradation level (same stress voltage and temperature) and subsequently, monitor threshold voltage recovery at temperatures differently from the stress temperature [5]. From these experiments several conclusions were drawn. First, interface states recovery is not significantly influenced by temperature and is most probably permanent. Second, there is a growing initial offset between recovery-curves recorded at different temperatures indicating a very fast (ms range) temperature dependent component. Third, after some seconds post stress the recovery rate per decade is roughly constant at all investigated temperatures making a diffusion controlled mechanism very unlikely. In short, these results again favor a tunneling rather than a diffusion component, since an irrational variation of diffusion barriers would be necessary in order to explain such a behavior by hydrogen diffusion [6-8]. In this

paper we extend our knowledge on the temperature dependence of NBTI recovery by making progressive use of the established poly heater technique [9-10]. By using the feature to vary temperature directly during recovery (while keeping the gate bias applied), we can rule out without any doubt elastic tunneling as well as interface states passivation to be the dominant processes in NBTI recovery.

To minimize direct tunneling currents and FN-Injection during recovery, we use PMOS transistors with 30nm SiO_2 gate oxides and n++ gate poly doping. It has been shown that the basic mechanisms behind NBTI are essentially the same for thick SiO_2 and SiON technologies [11].

II. In Situ Heating And Cooling

On the test chip single MOS devices are embedded between poly heater wires which can act like local radiators. If a certain electrical power is applied to the wires, the local temperature of the surrounded device quickly approaches a defined elevated temperature. The heater-power vs. device-temperature can be calibrated by using the saturation drain current around V_{TH} as a thermometer. The target drain currents for the initial calibration have to be determined before the actual measurement with the help of the thermo chuck. For that purpose we heat up the chuck (and thereby device junction and interface) to a proper temperature T_j and then record transfer characteristics of the unstressed device. For a defined gate voltage we extract drain current vs temperature tuples (Id_j,T_j). During the subsequent poly-heater calibration cycle at a lower chuck temperature T_i, the heater power is increased incrementally until the measured drain current corresponds again exactly to its target value at T_j. The output of the procedure is a well defined power (P_{ij}) that has to be applied to the heater in order to reach a certain gate oxide interface temperature (T_j) within the MOS device. During our experiments we use the calibrated heater powers P_{ij} as representatives for certain device temperatures T_j at basically arbitrary chuck temperatures $T_i \leq T_j$. Once all desired power vs temperature couples (P_{ij},T_j) have been collected, no further drain current measurement for power adjustment is necessary during the experiment. This would be impossible anyway since the drain current is known to degrade as a result of NBTI. To ensure a constant device temperature even during long-term measurement (>10s), a program internal regulation algorithm checks permanently the instantaneous power supply and corrects the heating voltage if a power decrease is detected due to poly degradation.

If we reduce or take away the heating power, the device cools down immediately (~1s). The lowest achievable temperature is defined by the temperature of the underlying thermo chuck (at zero heating power). The heating and cooling characteristics of the device as well as a schematic description of the poly heater layout are illustrated in Fig. (1). In our studies we use this feature to quickly vary the temperature during degradation and recovery without loosing needle contact (gate biasing). In the following measurements the thermo chuck was always held at -40°C and every temperature above -40°C was generated by the poly heater.

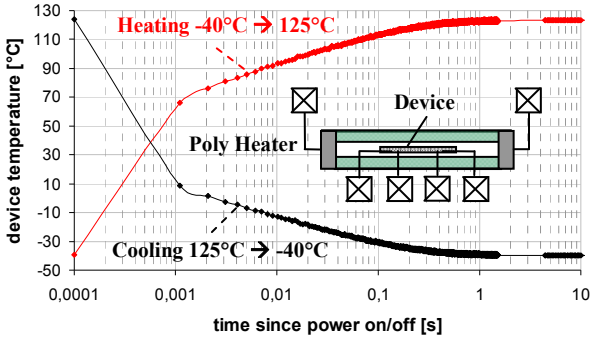

Fig (1): Cooling/Heating characteristics of the device as we turn off/on the heater. After about one second the device is nearly at target temperature. Heating and cooling durations are comparable. A schematic description of the in-situ heater is given in the inset.

III. Bias and Temperature Effects

The key experiment motivating our investigations is displayed in Fig. (2). We have stressed identically processed devices at same field and temperature conditions, and then varied the recovery temperature **and** included a short gate bias switch. This was to highlight the basic characteristics of NBTI recovery. The capability of stressing a device at temperature T_j and afterwards immediately letting it recover at temperature $T_i \neq T_j$ is a feature that can so far just be provided by the poly heater technique. Previous experiments [2-3] could not separate bias and temperature effects on recovery as they were limited to one temperature equal to stress temperature.

After stress of three identical devices at 5.5MV/cm and 125°C, we let them recover for 1.5s at a gate bias of -1.1V and at temperatures of -40°C/40°C/125°C. Within this 1.5s time interval the observed V_{TH} shifts look similar to the ones reported in [5]. After that first period of "standard" recovery we have changed the gate bias for 1.5s from -1.1V (deep inversion) to -0.3V (weak inversion), illustrated by the dashed box in Fig. (2). From a physical point of view the gate bias change corresponds to a temporary Fermi level shift from the valence band edge towards mid gap. As a consequence, the concentration and energies of the free substrate carriers differ from their respective values at threshold voltage. After switching the bias back to -1.1V, the V_{TH} shift is considerably reduced and the recovery curves are nearly flat for the following 1000s. **The remaining shift ΔV_{TH} is influenced both by temperature and gate bias**. The gate bias influence, by its mentioned impact on carrier energy and concentration, suggests a tunneling mechanism for the observed recovery in ΔV_{TH}. However, as temperature also plays a significant role, elastic tunneling from the substrate into the oxide can not explain the entire recovery effect. In the following the temperature influence is studied in more detail.

IV. Temperature Change During Recovery

We have used the poly heater feature in this section to study the effects of **heating and cooling during recovery**. In our recovery experiments we record the degradation of the saturation drain current (Id_{sat}) at constant biasing conditions and later convert variations in Id_{sat} into appropriate threshold voltage shifts [6]. It is essentially important here to take care on the intrinsic temperature dependence of Id_{sat}. A reliable

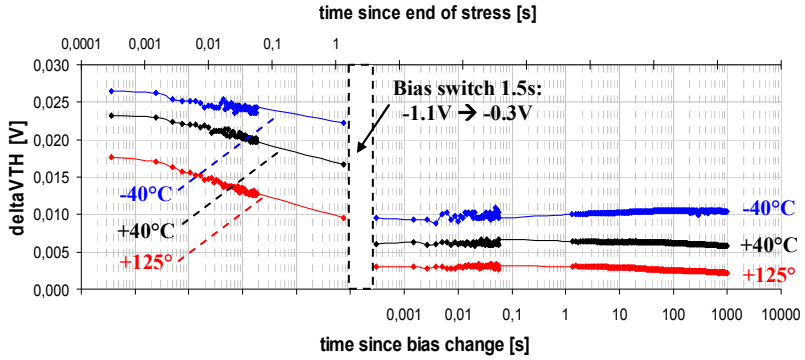

Fig (2): Influence of a short gate bias change (1.5s switch from -1.1V → -0.3V) on recovery. Illustrated are three different recovery temperatures ranging from -40°C to +125°C. All devices were stressed at 5.5MV/cm and 125°C. During the bias switch the Fermi level is shifted towards midgap which speeds up recovery for a certain time. As the bias comes back to V_{TH}, we observe accelerated recovery. The remaining degradation is the smaller, the higher the temperature.

conversion ΔId_{sat} ←→ ΔV_{TH} is just acceptable if we reference to the correct device temperature. Thus, transfer curves of the unstressed device are needed at every analysed temperature. However, if we want to evaluate threshold voltage recovery during temperature change, a problem arises due to the fact that the heater-device system needs a couple of seconds to restore thermal equilibrium (see Fig(1)). Within this short period of time the device temperature is more or less undefined. The calculation of ΔV_{TH} becomes erroneous if we reference the V_{TH} shift to the target temperature immediately after the power switch.

In Fig. (3) V_{TH} recovery of three identically processed devices is illustrated. All devices were stressed at 5.5MV/cm and 125°C for the same time. Two reference devices recovered at a constant temperature of -40°C and 125°C. The third device initially recovered at -40°C for 10s, and then at 125°C for another 10000s. As can be seen in Fig. (3), **at higher temperature recovery is accelerated** and approaches the 125°C reference curve after a couple of seconds. Furthermore we note a remarkable behaviour of ΔV_{TH} at the moment the temperature changes. This is due to the fact that the calculation of ΔV_{TH} within the short time interval of heating -40°C→125°C is afflicted with a systematic error. The error is originated in the intrinsic temperature dependence of the drain current and leads to a parasitic ΔV_{TH} component as long as the temperature is not stable at 125°C.

$$\Delta V_{TH}^{total}(t) = \Delta V_{TH}^{\Delta T}(t) + \Delta V_{TH}^{str/rec}(t) \qquad (1)$$

In equation (1) ΔVTH_total(t) is the extracted threshold voltage shift without any correction, ΔVTH_ΔT(t) is the parasitic component in the undefined temperature regime between -40°C and 125°C and ΔVTH_str/rec(t) is the stress and recovery induced threshold voltage shift. The component ΔVTH_ΔT(t) is completely independent of stress and becomes zero as the device reaches its constant target temperature. Since we are only interested in the ΔVTH_str/rec(t) term, we had to find a way to eliminate the systematic error during heating. This can be done by performing the same experiment in advance without stressing the device by gate bias. Since such an experiment does not lead to any degradation the factor ΔVTH_str/rec(t) is zero. What remains is the desired correction term ΔVTH_ΔT(t). The output of the reference measurement (without gate stress) is illustrated in Fig. (4). The corrected final graph ΔVTH_str/rec(t) can be obtained by subtracting Fig. (4) by Fig(3) and is illustrated in Fig(5).

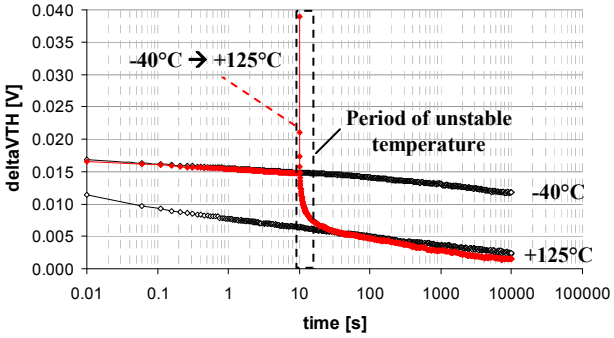

Fig (3): Temperature accelerated recovery. The straight curves are reference measurements at -40°C and 125°C without temperature switch. Due to heating from -40°C to 125°C, recovery is accelerated remarkably. As long as the temperature is not stable, the calculation of ΔV_{TH} is afflicted with a systematic error.

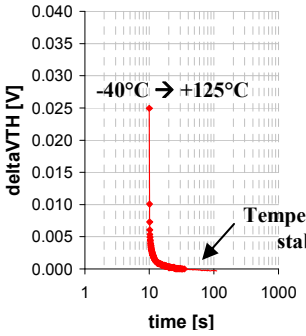

Fig (4): Reference measurement of deltaV$_{TH}$ without stressing the device. The parasitic V_{TH} shift is due to the finite time of heating and can be used to correct the systematic error in Fig. (3).

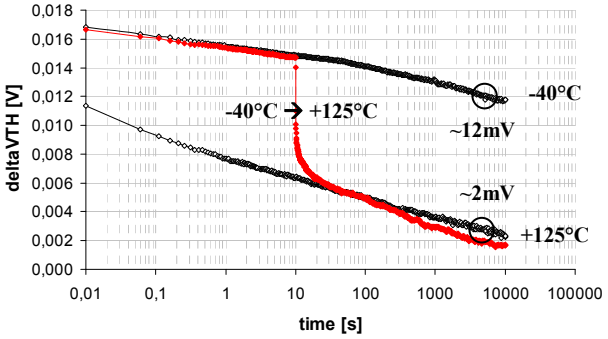

Fig (5): Temperature dependent recovery acceleration corrected by the systematic error illustrated in Fig. (4). At the end of recovery the difference in ΔV_{TH} between -40°C and 125°C is about 10mV.

Such a temperature accelerated recovery at constant bias conditions can definitely not be ascribed to elastic tunnelling. In order to check if this can be explained by interface states passivation, we have measured CP (Charge Pumping) [13-14] currents before stress and after stress + recovery for all three devices under test. It is essential to note that in this experiment all CP currents could be recorded at -40°C independently of the recovery temperature before, because the poly heater generated both the stress and recovery temperature (chuck always at -40°C). The short cooling duration from 125°C to -40°C (<10s) can safely be neglected in comparison to the 10000s recovery phase before. This is an innovative extension of the method described in [5] since it does not require weight factors or any assumptions on the density of states in order to compare CP currents of devices recovered at different temperatures.

As can be seen in Fig. (6), there is indeed a slight discrepancy in CP current whether we let the device recover at -40°C or 125°C. However, when converting the differential CP signal into an effective threshold voltage shift, we obtain values $\Delta VTH\_it < 1$ mV that can neither explain the finally observed ΔV_{TH} of 12 and 2 mV for -40°C and 125°C, respectively, nor the difference of about 10 mV between the two. The conversion $\Delta ICP \leftrightarrow \Delta VTH\_it$ is explained in the appendix. We conclude that **interface states do not play a crucial role in recovery and final degradation**.

In the experiment illustrated in Fig. (7), we have elevated the temperature twice (-40°C→40°C→125°C) during recovery. Similarly to Fig. (5), the recovery curve approaches the appropriate reference curves after a couple of seconds.

In Fig. (8), we have performed the complementary experiment to Fig. (5): at first, we let the device recover at 80°C for 10 s, and then lowered the heater power so that the device cools down to 40°C. As can be seen, cooling leads to frozen recovery for a certain time interval. Recovery does not proceed before the cooled measurement curve reaches the reference curve at 40°C.

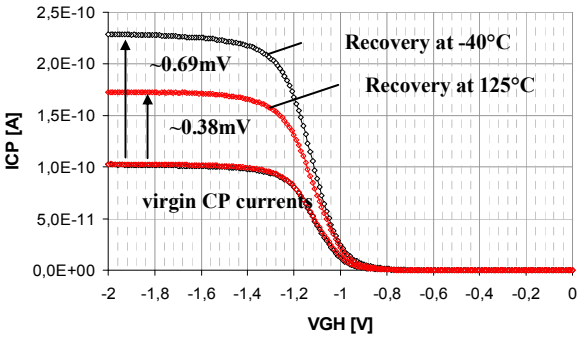

Fig (6): CP currents recorded before stress and after 10000s recovery at -40°C and 125°C (cf. Fig. (5)). Although heating seems to accelerate interface states recovery slightly, the small offset cannot be made responsible for the huge V_{TH} difference in Fig. (5).

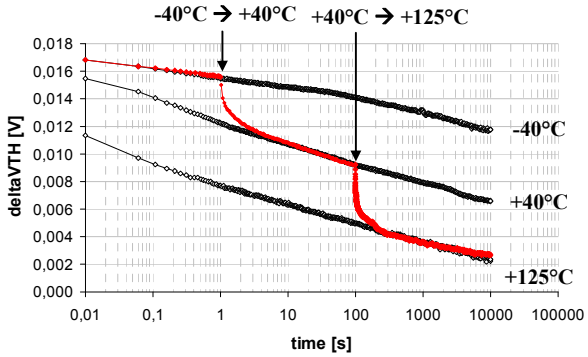

Fig (7): Two step heating during recovery. Reference measurements at -40°C/+40°C/125°C are illustrated. Recovery can be accelerated twice as we heat up to +40°C and 125°C.

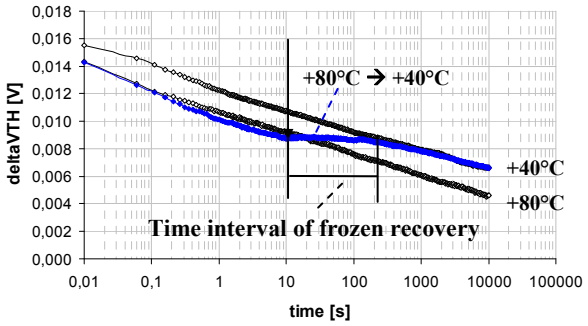

Fig (8): Cooling during recovery has the opposite effect as heating. As a consequence to lowering the recovery temperature, the V_{TH} recovery seems to be frozen for a certain time interval. Not until the cooled curve reaches the reference curve at +40°C recovery proceeds.

V. DISCUSSION

Since our measurements show both bias and temperature dependence, but support neither elastic tunneling nor interface states passivation, a different mechanism has to be responsible for recovery via carrier trapping. As opposed to elastic tunneling, inelastic phonon assisted tunneling is temperature dependent [16]. Oxide defects and valence band electrons having different energetic positions cannot exchange carriers elastically. However, if they gain energy from lattice vibration (phonons) they can pass the thermo dynamical tunneling barrier ΔE_B [17-18] with a certain temperature dependent probability. The lifetime of such a trap can be expressed by an Arrhenius law

$$\tau(\Delta E_B, T) = \tau_0 \exp\left(\frac{\Delta E_B}{kT}\right) \qquad (2)$$

In equation (2), $\tau(\Delta E_B,T)$ is the inelastic tunneling lifetime of a trap, τ_0 is the elastic tunneling exchange time between trap and substrate carrier at barrier height zero, ΔE_B is the thermo dynamical tunneling barrier, k the Boltzmann constant and T the temperature. At a constant temperature T the time constants of different traps are solely determined by their hypothetical barrier height ΔE_B.

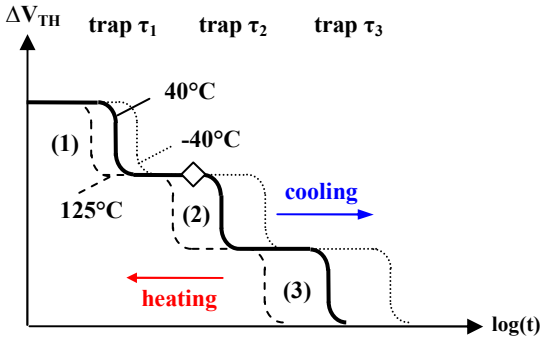

Fig (9): Schematic explanation of the recovery of three different traps via inelastic phonon assisted tunnelling. Heating or cooling shifts the recovery curve to the left or the right (shorter or longer time constants) resulting in stimulated or decelerated recovery. The diamond indicates an imagined temperature switch.

If we consider NBTI recovery to be mainly determined by the neutralization of positive oxide defects via electron capture from the silicon substrate (respectively hole emission into the silicon substrate), the observed threshold voltage recovery can be interpreted as a continuous decay of traps with different barrier heights ΔE_B. Based on this idea, we can illustrate the ΔV_{TH} recovery curve for three traps schematically in Fig. (9):

(i) Due to different barrier heights, each trap will recover at a certain characteristic time (ΔV_{TH} plateau).

(ii) A variation of temperature ($T_1 \rightarrow T_2$) modifies all time constants in parallel and shifts the plateaus parallel along the time axis in log scale. The respective shift in time for a trap with barrier height ΔE_B can be denoted as

$$\log \tau(\Delta E_B, T_1) - \log \tau(\Delta E_B T_2) = \frac{\Delta E_B}{kT_1} - \frac{\Delta E_B}{kT_2} \qquad (3)$$

For $T_2 > T_1$, the plateaus will shift to the left since all time constants of all traps will decrease, for $T_2 < T_1$ the time constants increase leading to a shift to the right. We further note that the widths of the plateaus are proportional to ΔE_B. The trap level, which recovers at first (τ_1), has the lowest barrier (ΔE_{B1}) and is therefore least temperature dependent. On the other hand trap levels with higher barriers (τ_2 and τ_3) are stronger dependent on T which results in a more significant temperature impact on the plateau broadness.

In this model, heating up or cooling down the device **during** recovery leads to temperature acceleration (at the diamond, stepping from the solid to the dashed line in Fig. (9), compare with Fig. (5)) or freezing (at the diamond, stepping from the solid to the dotted line in Fig. (9), compare with Fig. (8)) like proposed by our measurement results. If we further assume that the tunneling barrier ΔE_B itself can be lowered by a bias change, the model covers also bias change experiments and includes elastic tunneling in the limit $\Delta E_B = 0$.

VI. Conclusion

Bias and temperature dependent recovery measurements of identically stressed devices unambiguously demonstrate that neither elastic tunneling nor interface states passivation can be the dominant process in NBTI recovery. Alternatively, our data strongly suggest oxide trap passivation by inelastic phonon assisted tunneling. What we interpret as recovery acceleration could be originated in carrier exchange processes between the silicon substrate and NBTI induced oxide traps. Continuously distributed thermo dynamical tunneling barriers would lead to a large variety of time constants resulting in a numerous number of small single steps like the ones described in Fig. (9). In a realistic experiment (large device) we would therefore expect that large amount of small steps to be smeared out as a straight line in a log(t) diagram.

Appendix

The conversion $\Delta ICP \leftrightarrow \Delta VTH\_it$ in Fig. (6) was done under the assumption of an amphotheric nature of interface traps [15]. This means, interface traps above the intrinsic Fermi level were considered as acceptor like while traps below E_i were treated as donor like. In thermal equilibrium traps above the Fermi level E_f are empty while traps below E_f are occupied by an electron. This means that only traps between E_f and E_i can contribute effectively to a positive V_{TH} shift at threshold voltage. The position of E_f at a gate voltage of -1.1V was calculated to be between 100 meV (-40°C) and 200 meV (125°C) above the valence band edge for our PMOS devices. Following [14] we can calculate the emission boundaries for the maximum CP current at -40°C as well. Using our pulse setup and under the assumption of a energetically homogeneous capture cross section of 10^{-15} cm^2 we found the emission boundaries for CP to be approximately 200 meV away from the silicon band edges. Obviously in this experiment the contributing energy interval in CP was about half of the electrically active energy interval at a constant gate voltage around threshold. This factor of two comes from the symmetrical CP pulse setup. If we use the same rising and falling slopes for the gate pulse the upper and lower half of the silicon band gap is scanned in parallel. Due to the coinciding active energy intervals we can calculate straight forward the interface states dependent threshold voltage shift for our recovery experiment,

$$\Delta V_{TH}^{it} \approx \frac{\Delta Q_{it}}{C_{OX}} = \frac{\Delta I_{CP}^{\max}}{2 \cdot f \cdot A \cdot C_{OX}} \qquad (A1)$$

In equation (A1) ΔVTH_it is the positive threshold voltage shift due to the build up and charging of interface states, ΔQ_{it} is the increase in electrically active traps, C_{OX} is the area-related oxide capacitance, A is the device area, ΔICP_max is the increase in CP current and f is the CP measurement frequency. Insertion of the appropriate experimental and device parameters into equation (A1) yields the values for ΔV_{TH} quoted in Fig. (6).

Acknowledgment

The authors gratefully acknowledge support in the polyheater design and layouting process by Sascha Baier, Norbert Krischke and Tom Ostermann (Infineon Technologies).

This work was jointly funded by the Federal Ministry of Economics and Labour of the Republic of Austria (contract 98.362/0112-C1/10/2005) and the Carinthian Economic Promotion Fund (KWF) (contract 98.362/0112-C1/10/2005).

REFERENCES

[1] D. K. Schroder, "Negative bias temperature instability: what do we understand?", Microelectronics Reliability, Vol. 47, 2007, pp. 841-852.

[2] V. Huard, M. Denais and C. Parthasarathy, "NBTI degradation: from physical mechanism to modeling," Microelectronics Reliability, Vol. 46, 2006, pp. 1–23.

[3] T. L. Tewksbury, Hae-Seung Lee, "Characterization, Modeling, and Minimization of Transient Threshold Voltage Shifts in MOSFET's", IEEE Journal of solid-state circuits, Vol. 29, No. 3, 1994, pp. 239-252.

[4] M. A. Alam, H. Kufluoglu, D. Varghese, S. Mahapatra, "A comprehensive model for PMOS NBTI degradation: Recent progress", Microelectronics Reliability, Vol. 47, Issue 6, 2007, pp. 853-862.

[5] T. Aichinger, M. Nelhiebel, T. Grasser, "On the temperature dependence of NBTI recovery", Microelectronics Reliability, Vol. 48, Issue 8-9, 2008, pp. 1178-1184.

[6] 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", IRPS 2005, pp. 381- 387.

[7] T. Grasser, W. Goes, B. Kaczer, "An energy-level perspective of bias temperature instability", IRPS 2008, pp. 28-38

[8] R. Biswas, Y.-P. Li, B. C. Pan, "Enhanced stability of deuterium in silicon", Applied Physical Letters, Vol. 72, No. 26, 1998, pp. 3500-3502.

[9] R. Biswas, Qiming Li, B. C. Pan, Y. Yoon, "Mechanism for hydrogen diffusion in amorphous silicon", Physical Review B, Vol. 57, No. 4, 1998, pp. 2253-2256.

[10] W. Muth, W. Walter, "Bias temperature instability assessment of n- and p-channel MOS transistors using a polysilicon resistive heated scribe lane test structure", Microelectronics Reliability, Vol. 44, Issue 8, 2004, pp. 1251-1262.

[11] C. Schluender, M. Hoffmann, R.-P. Vollertsen, G. Schindler, W. Heinrigs, W. Gustin, H. Reisinger, "A new smart Vth-extraction methodology considering recovery and mobility degradation due to NBTI", IIRW Final Report, 2007, pp. 1-5.

[12] T. Grasser, W. Goes, B. Kaczer, "Dispersive Transport and Negative Bias Temperature Instability: Boundary Conditions, Initial Conditions, and Transport Models", IEEE Transactions on Device and Materials Reliability, Vol. 8, Issue 1, 2008, pp. 79-97.

[13] J. S. Brugler, P. G. A. Jespers, "Charge pumping in MOS devices", IEEE Transactions on Electron Devices, Vol. 16, Issue 3, 1969, pp. 297-302.

[14] G. Groeseneken, H. E. Maes, N. Beltran, R. F. De Keersmaecker, "A Reliable Approach to Charge-Pumping Measurements in MOS Transistors", IEEE Transactions on Electron Devices, Vol. 31, No. 1, 1984, pp. 42-53.

[15] J. Bardeen, "Surface States and Rectification at a Metal Semiconductor Contact", Phys. Rev. B, Vol. 71, 1947, pp. 717-727.

[16] S. D. Ganichev, E. Ziemann, W. Prettl, "Distinction between the Poole-Frenkel and tunneling models of electric-field-stimulated carrier emission from deep levels in semiconductors", Phys. Rev. B Vol. 61, No. 15, 2000, pp. 61-65.

[17] T. Grasser, B. Kaczer, W. Goes, T. Aichinger, P. Hehenberger, M. Nelhiebel, "A Two-StageModel for Negative Bias Temperature Instability", unpublished (IRPS 2009)

[18] W. Goes, M. Karner, S. Tyaginov, P. Hehenberger, T. Grasser, "Level shifts and gate interfaces as vital ingredients in modeling of charge trapping", SISPAD 2008, pp. 69-72