Impact of Hydrogen on Recoverable and Permanent Damage following Negative Bias Temperature Stress

T. Aichinger¹, S. Puchner¹, M. Nelhiebel², T. Grasser³, and H. Hutter⁴

¹Kompetenzzentrum für Automobil- und Industrieelektronik (KAI), Europastraße 8, 9524 Villach, Austria
phone: 0043-4242-34890-24, email: thomas.aichinger@k-ai.at
²Infineon Technologies Austria, Siemensstrasse 2, 9500 Villach, Austria
³Christian Doppler Laboratory for TCAD, Institute for Microelectronics TU Wien, Gusshausstrasse 27-29, 1040 Wien, Austria
⁴Institute for Chemical Technologies and Analytics TU Wien, Getreidemarkt 9/E164, 1060 Wien, Austria

Abstract—By subjecting selected split wafers to a specifically adapted measure-stress-measure (MSM) procedure, we analyze negative bias temperature stress (NBTS) and recovery characteristics of PMOS devices with respect to the impact of hydrogen. We control the hydrogen incorporation within the gate oxide during Back End of Line (BEOL) fabrication by varying the titanium barrier thickness below the routing metallization. Differences in the initial passivation degree of the gate oxide are verified electrically by Charge Pumping (CP) measurements and physically by time-of-flight secondary ion mass spectrometry (TOFSIMS). Our results indicate that the total $V_{TH}$ shift is the sum of quasi permanent and recoverable damage which are of comparable scale but have completely different physical and electrical characteristics. While the permanent component seems to be strongly linked to hydrogen release from the interface (increase in CP current), the recoverable component is widely independent of hydrogen and its recovery can be controlled via carrier exchange with the silicon substrate. Hence, our results suggest different trap precursors for the individual components which challenge some predictions of the classical reaction-diffusion (RD) model and support the concepts of an alternative model based on permanent interface state creation via hydrogen transfer to recoverable $E'$ centers which have their origin in oxygen vacancies, whose density is roughly independent of the hydrogen concentration.

I. INTRODUCTION

Hydrogen has often been reported to play a crucial role in the negative bias temperature instability (NBTI) of MOS devices [1-4]. This has been linked to its ability to passivate/de-passivate dangling bonds at the gate-oxide silicon-substrate interface [5]. In particular, the performance and defect densities of virgin silicon devices improve considerably by incorporating hydrogen into the gate oxide during the Back End of Line (BEOL) process [6]. This can be done either directly by exposing the wafers to pure hydrogen or forming gas anneals [7-8] or indirectly as a consequence of plasma-enhanced chemical vapor deposition (PECVD) of silicon nitride (SNIT) layers [9]. Such layers are considered to be efficient hydrogen sources since they contain a large concentration of hydrogen [10-11] which may be released immediately after deposition to diffuse toward the gate oxide. Provided there is no diffusing barrier below the SNIT, some hydrogen may reach the gate oxide where it can passivate dangling bonds at the interface, thereby improving the virgin performance of the MOS device. However, once passivated, previously captured hydrogen may be released from the interface during NBTS leaving behind donor-like $P_b$ centers that are reported to cause a negative shift in the threshold voltage of a p-doped metal oxide semiconductor (PMOS) transistor. Thus, one may expect that the initial passivation degree of the interface, namely the total number of Si-H bonds present at the interface before stress, crucially determines the NBTI sensitivity of the technology. This is a generally accepted fact often reported in literature. However, concerning the underlying micro-structural physics behind degradation and recovery, two competing models (namely the reaction diffusion (RD) model [2] and the alternative Grasser model [12]) come to different conclusions and hence make different predictions.

In order to check the fundamental statements of those models with respect to the impact of hydrogen, we have fabricated a wafer split, where we modified the hydrogen budget within the gate oxide (and hence the number of Si-H bonds at the interface). The hydrogen incorporation is measured physically by TOFSIMS analysis (counting secondary H ions) and electrically by CP measurements (counting dangling bonds at the interface). Having fabricated the hardware, we characterize and compare stress and recovery dynamics of two selected split wafers. Finally, we check our results against the predictions of the classical RD model [2] and against the alternative model introduced by Grasser et al. [12].

II. SAMPLE PREPARATION AND HARDWARE

Our basic idea for modifying the hydrogen content within the gate oxide is to control hydrogen diffusion from the upper SNIT layer toward the gate oxide. This is basically achieved by introducing titanium (Ti) liners below the routing metallization. Titanium is known to be an effective hydrogen barrier [13-15] which can absorb a certain amount of diffusing hydrogen.

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).
in-situ polyheaters that enable us to perform on-chip fast heating and cooling [16-17]. We use such thick oxide devices in order to improve the hydrogen resolution within the gate oxide during the TOFSIMS measurement and in order to reduce tunneling currents through the gate oxide during stress and during charge pumping. It has recently been shown that the basic mechanisms of NBTI are essentially the same in thin and thick SiO2 and SiON technologies [18-20]. The use of pure SiO2 gate oxides guarantees that our general conclusions are not distorted by the strongly process dependent impact of nitridation [21].

III. HYDROGEN DETECTION USING TOFSIMS

The hydrogen concentrations of the different split wafers were investigated by a time of flight secondary ion mass spectrometer (TOF.SIMS\(^3\)) equipped with a liquid metal ion gun (LMIG) producing primary ions and a dual source (DSC-S) column eroding the sample surface. TOFSIMS measurements were performed on large reference capacitors (1 mm\(^2\)) located in the vicinity of the electrically measured PMOS devices. Depth profiling was carried out in the dual beam mode [22] using common techniques for insulators [23]. Pulsed Bi\(^+\) primary ions (25 keV) were used to analyze the target surface. Alternating, a second beam of Cs\(^+\) ions erodes the sample surface and enhances the yield of negatively charged secondary ions (e.g. H\(^+\), O\(^-\)). In order to avoid distortion of the hydrogen signal, we abstained from exposing our samples to a chemical back preparation via etching. Hence, at the expense of depth resolution, we were forced to use a 2 keV sputter beam ablating the thick stack of layers (3.5 µm) on the top of the gate oxide within a reasonable time interval. To quantify the hydrogen signal, a relative sensitivity factor (RSF) was used for converting the secondary ion intensity into absolute concentrations. For determining this hydrogen-related RSF, we used reference samples providing a defined standard implantation dose (\(\Phi_H = 10^{15}\) cm\(^2\)). The obtained RSF for hydrogen was 5x10\(^{-21}\), which is close to literature values taking into account different matrix signals and SIMS techniques [24].

Fig. 1 (a) shows a TOFSIMS image of the sub-metal BEOL layer stack of the two selected split wafers. Displayed are the oxygen and the titanium signals for the orientation within the BEOL stack. Underneath the Ti liners we measure significantly different hydrogen concentrations in the PMD, the poly and the GOX for wafer I and II. In perfect agreement with the TOFSIMS results we obtain in Fig. 1 (b) that (i) the initial CP signal of wafer I (~ 0.3 nA) is about 30 times lower than the one of wafer II (~ 9 nA). This is consistent with the assumption that the interface of wafer I (Dit \(\sim 2.3 \times 10^{17}\) eV\(^{-1}\)cm\(^{-2}\)) is more efficiently passivated by hydrogen than the interface of wafer II (Dit \(\sim 6.9 \times 10^{16}\) eV\(^{-1}\)cm\(^{-2}\)). The virgin CP currents were recorded at a temperature of 50°C using a pulsing frequency of 500 kHz and rising/falling slopes of 10 V/µs scanning roughly 500 meV of the silicon band gap around mid gap; (ii) wafer II has a more negative threshold voltage (~ -160 mV) than wafer I which can be explained by an initially larger number of positively charged Pb centers and oxide defects which both remain unpassivated in wafer II due to the low hydrogen concentration within the gate oxide.

To produce split wafers with different hydrogen concentrations within the gate oxide, we simply vary the Ti liner thickness between the SNIT and the gate oxide. The thinner the titanium liner, the higher is its hydrogen permeability and the more hydrogen arrives actually at the gate oxide. In particular, as the titanium layer thickness exceeds several tens of nanometers, the hydrogen concentration within the gate oxide approaches a minimum which is reflected electrically by a maximum initial CP signal (low interface state passivation degree). In fact, the hydrogen concentration within the gate oxide approaches a minimum which is reflected electrically by a maximum initial CP signal (low interface state passivation degree). In our measurements discussed in the following, we focus on two representative split wafers (extreme cases), the first (wafer I) having a minimum titanium barrier (high hydrogen concentration in the gate oxide), the second (wafer II) having a very thick titanium barrier (minimum hydrogen concentration in the gate oxide). Except for the Ti-liner thickness, the two wafers were processed identically in the same wafer lot. Our devices under test (DUTs) are isolated PMOS transistors with 30 nm pure SiO2 gate oxides. The structures are surrounded by
IV. QUALITATIVE MODEL DISCUSSIONS

Before starting with the description of the measurement setup and presenting results, we briefly summarize some of the main characteristics of the discussed NBTI models.

A. The RD model:

The classic reaction-diffusion (RD) model as described in [2] explains $V_{TH}$ degradation and recovery by a non-dispersive diffusion process of neutral hydrogen atoms or molecules (H or $H_2$) which are released during stress from Si-H bonds at the interface via a field-dependent reaction. Once released, the hydrogen species diffuse away from the interface, leaving behind a positively charged interface state (Si$^+$) that is responsible for the higher threshold voltage and the lower transconductance. Once the stress is removed, previously released hydrogen can diffuse back and recombine with silicon dangling bonds restoring them to their passive Si-H state.

Consequences: (i) the increase in CP current fully explains the total $V_{TH}$ shift; (ii) the recovery rate is proportional to the amount of created interface traps, released hydrogen respectively; (iii) recovery is not influenced by slight variations in gate biasing during read-out.

B. The Grasser model:

The chemical transitions suggested by the reaction controlled model of Grasser et al. [12] are schematically depicted in Fig. 2. Path A: During NBTS, oxygen vacancies located close to the interface are assumed to break up and charge positively (transition I) due to the presence of the high electric field and a majority of holes at the gate oxide substrate interface. During recovery, where the field and the carrier situation at the interface is quite different, some of these so-called E' centers ($Q_{ox}^{rec}$) may become neutralized by hole emission (transition II). Once in the neutral charge state, the E' center and travels back to the un-passivated interface state, is feasible as well. However, in a first order approximation, this back transition is neglected assuming that the Si-H bond is stable within the E' center. Consequently, once created, locked-in oxide defects and interface states are considered as permanent.

Consequences: (i) the total $V_{TH}$ degradation consists of permanent interface states and locked-in oxide charges (which emerge in a physical 50:50 relation due to entropy driven hydrogen exchange between Si-H bonds at the interface and positively charged E' centers (cf. Path B)) plus a recoverable portion of positively charged E' centers; (ii) recovery (cf. Path A) is largely independent of the hydrogen incorporation (at least, when neglecting the loss of recoverable E' centers by transition IV); (iii) recovery (cf. Path A) is highly dependent on the gate bias during read-out since the carrier situation at the interface (and hence the Fermi-level position) governs neutralization and relaxation of positively charged E' centers.

V. ELECTRICAL MEASUREMENT SETUP

Our experimental setup for NBTI characterization is illustrated in Fig. 3. The measurement procedure is particularly designed to separate time and bias dependent recoverable damage from apparently permanent degradation. During a basic measure-stress-measure (MSM) cycle we subject PMOS devices to NBTS ($200°C < 7.0 \text{ MV/cm}$) for a defined time $t_s$. By making use of the in-situ polyheater technique, the following recovery phase can be performed at a much lower temperature of 50°C which decelerates thermo dynamical recovery mechanisms and improves the charge pumping measurement resolution. Right before the end of the stress phase, we quickly (< 10 s) cool down to 50°C while the stress bias is maintained at the gate, thereby quenching the degradation [16-17]. Then we initiate a 1000 s recovery phase (tR1), by switching the gate bias from its stress level to -2.0 V.

![Figure 2. A schematic illustration of the Grasser model transitions.](image)

![Figure 3. Our basic MSM procedure used for degradation/recovery analysis.](image)
We denote the relative amount of $V_{TH}$ recovery during $t_{R1}$ between the first measured point after the removal of the stress bias (40 ms post stress) and the last measured point (1000 s post stress) as the time dependent recovery contribution ($\Delta V_{TH,\text{time}}$). Subsequently to $t_{R1}$, we ramp down (SD1) the gate bias in 20 mV steps from strong inversion (-2.0 V) toward depletion (0.0 V) and monitor in parallel the $V_{TH}$ shift as a function of the current gate bias. One full gate bias ramp takes approximately 10 s. Approaching depletion, the Fermi level moves from the valance band edge toward the conduction band edge thereby gradually changing the ratio of free holes and electrons at the interface. After staying for 10 s at 0.0 V, we ramp the gate bias back (SU1) to -2.0 V. The difference in the $V_{TH}$ shift recorded at -2.0 V at the beginning of SD1 and at the end of SU1 is denoted as the bias dependent recovery contribution ($\Delta V_{TH,\text{bias}}$). After the first ramp down-up cycle, we record the maximum CP current for 10 s by pulsing the gate voltage in 20 mV steps from strong inversion (-2.0 V) toward depletion (0.0 V) at a frequency of 500 kHz. In the analysis, we convert the maximum CP signal into an interface state dependent VTH shift and the interface state dependent VTH shift are much larger for wafer I than for wafer II, cf. Fig. 4 (c2); (iv) after the intermediate CP cycle, the remaining degradation level is permanent and cannot be reduced further by an additional gate bias ramp toward 0.0 V; (v) The remaining permanent $V_{TH}$ shift and the interface state dependent $V_{TH}$ shift are much larger for wafer I than for wafer II, cf. Fig. 4 (c3); (vi) the interface state dependent $V_{TH}$ shift is smaller than the permanent $V_{TH}$ shift, cf. Fig. 4 (c3); (vii) interface state dependent and permanent $V_{TH}$ shift coincide when multiplying $\Delta V_{TH,\text{perm}}$ by a factor 3, cf. Fig. 4 (c4).

![Diagram](https://via.placeholder.com/150)

**Figure 4.** The $V_{TH}$ shifts recorded after six subsequent stress-runs (1/10/100/1,000/10,000/100,000 s) at 50°C at different stages of the experiment (wafer I (a); wafer II (b)). The time dependent $V_{TH}$ recovery ($\Delta V_{TH,\text{time}}$) is recorded at -2.0 V directly post stress for 1000 s (IR1). The bias dependent recovery ($\Delta V_{TH,\text{bias}}$) is the difference in $V_{TH}$ shift between SD1 and SU1 recorded at -2.0 V. Subsequently to SU1, we measure the maximum CP current and convert changes in $I_{CP,max}$ into appropriate interface state dependent $V_{TH}$ shifts ($\Delta V_{TH,\text{perm}}$). After gate pulsing, the remaining $V_{TH}$ shift at -2.0 V is permanent ($\Delta V_{TH,\text{perm}}$). In (c) the respective $V_{TH}$ shifts are summarized for both wafers as a function of stress time: (c1) $\Delta V_{TH,\text{perm}}$; (c2) $\Delta V_{TH,\text{bias}}$; (c3) $\Delta V_{TH,\text{perm}}$ (full symbols) & $\Delta V_{TH,\text{perm}}$ (open symbols).

By multiplying $\Delta V_{TH,\text{perm}}$ by a scaling factor 3, $\Delta V_{TH,\text{perm}}$ and $\Delta V_{TH,\text{perm}}$ coincide, cf. (c4).
These seven findings on the bias and time dependence of the recovery, on interface state creation and on permanent damage, as a function of the H content within the gate oxide, represent a significant collection of NBTI characteristics challenging the reliability of suggested models.

**Cross check with the proposed NBTI models:**

**Point (i) –** Recovery over time is H independent: Agrees only with the Grasser model which predicts that recovery (path A) is nearly independent of the hydrogen budget within the gate oxide except for the small fraction of recoverable damage being converted to locked-in oxide charge by hydrogen capture from the interface (path B), cf. Fig. 4 (c1). The RD model is challenged since it suggests a balance between hydrogen release and re-capture and would therefore predict a much larger amount of recovery for wafer I.

**Point (ii) –** $V_{TH}$ shift is dependent on the applied gate bias: Agrees with both models which suggest that oxide traps and/or interface states may become neutralized and charge up positively again as the Fermi level crosses the silicon band gap [27-28].

**Point (iii) –** Ramping the gate toward 0.0 V accelerates $V_{TH}$ recovery, the effect being H independent: Agrees only with the Grasser model which suggests bias dependent neutralization and relaxation of positively charged hydrogen-independent E' centers. The RD model cannot explain bias accelerated trap recovery since it assumes the hydrogen diffusion species to be neutral.

**Point (iv) –** The degradation level is quasi permanent after the intermediate CP cycle: In the Grasser model, this is explained by the bias-accelerated relaxation of E’ centers which is finished after the first gate bias double ramp. The RD model would not expect significant recovery within the following 10-100 s either, since already more than 1000 s elapsed since the actual end of stress.

**Point (v) –** The remaining permanent $V_{TH}$ shift is larger for the H rich wafer: Agrees with both models considering that the larger permanent damage of the H-rich wafer I is a logical consequence of its initially higher Si-H precursor concentration.

**Point (vi):** $\Delta V_{TH}^{perm,it}$ is smaller than $\Delta V_{TH}^{perm}$. Provided that our conversion of the CP signal into an interface state dependent $V_{TH}$ shift ($\Delta V_{TH}^{perm,it}$) is correct, statement (vi) agrees only with the Grasser model which suggests that the permanent $V_{TH}$ shift is the sum of locked-in oxide defects and interface states ($Q_{ox}^{perm} + Q_{it}^{perm}$). The RD model attributes the entire $V_{TH}$ shift solely to interface states.

**Point (vii) –** $\Delta V_{TH}^{perm,it}$ is proportional to $\Delta V_{TH}^{perm}$. Agrees with the Grasser model which predicts the simultaneous creation of $\Delta V_{TH}^{perm,ix}$ and $\Delta V_{TH}^{perm,ix}$ via hydrogen exchange between passivated interface states and E’ centers and hence suggests a physical 50:50 relation and a strong correlation between $\Delta V_{TH}^{perm,it}$ and $\Delta V_{TH}^{perm}$. The deviation in the measured factor 3 from the proposed factor 2 (50:50) may be due to different energy distributions of interface and oxide charges, leading to a different electrical response of both trap types. It has to be remarked that basically the profiled energy range during CP and the energy range of defects being positively charged during the drain current measurement do not coincide. We have considered this energy mismatch assuming an amphoteric nature of interface traps [25] and a flat density of state profile [26].

**Discussion on the absolute degradation potential**

Fig. 5 illustrates the development of the absolute CP current (a) and the effective difference in the absolute $V_{TH}$ between wafer I and II (b). Before stress (stress time = 0 s) wafer II has a considerable higher CP signal and also a much lower $V_{TH}$ than wafer I. Note that this initial difference in the CP current (~ 92 nA) accounts only for ~ 53 mV of the difference in the initial $V_{TH}$ (~ 160 mV) which reflects even on the virgin device the factor 3 measured between interface state creation and $V_{TH}$ shift. The mismatch is increasingly compensated with increasing stress time. Indeed, after 100,000 s of stress, both devices have a similar CP signal and almost the same $V_{TH}$. However, since neither the CP signal nor the $V_{TH}$ development of wafer I tends to saturate, it is likely to assume that wafer I would even have exceeded wafer II in the CP current and in the $V_{TH}$ at longer stress times. Following Ref. [6], this would indicate that the sum of Si-H bonds plus interface traps (Si•) is initially higher in a thoroughly passivated gate oxide, implying an additional latent damage by excessive hydrogen ingress. This may lead eventually to a higher absolute drift potential for the H-rich wafer I at longer stress times. Additional measurements are required in order to verify this suggestion unambiguously.

![Figure 5](image-url)

*Figure 5. Development of the maximum CP current as a function of the stress time for wafer I and II (a). After 100,000 s of stress, the maximum CP signal of wafer I equals the maximal CP signal of wafer II indicating a similar number of interface traps at the end of the last stress run. Also, the initial difference in $V_{TH}$ between wafer I and II is compensated at the end of the last stress run (b) indicating a similar degradation level of the oxide and the interface.*
VII. CONCLUSIONS

By varying the Ti thickness during BEOL processing, we have produced two split wafers with vastly different hydrogen contents within the gate oxide and subjected them to NBTS. In agreement with the Grasser model, we have demonstrated that the recoverable part of the NBTI degradation is largely independent of hydrogen while the permanent $V_{TH}$ shift component is strongly linked to the total hydrogen budget within the gate oxide. Also, a strong correlation between the increase in the CP signal and the permanent $V_{TH}$ shift component was found which is also consistent with the Grasser model.

REFERENCES


