Microelectronics Reliability 49 (2009) 1013–1017

Contents lists available at ScienceDirect

Microelectronics Reliability

journal homepage: www.elsevier.com/locate/microrel

On the temperature and voltage dependence of short-term negative bias temperature stress

Ph. Hehenberger [a,*], P.-J. Wagner [b], H. Reisinger [c], T. Grasser [b]

[a] *Institute for Microelectronics, TU Wien, Gußhausstraße 27–29, A-1040 Wien, Austria*
[b] *Christian Doppler Laboratory for TCAD at the Institute for Microelectronics, Gußhausstraße 27–29, TU Wien, A-1040 Wien, Austria*
[c] *Infineon Technologies, D-81739 München, Germany*

ARTICLE INFO

Article history:
Received 19 June 2009
Available online 7 August 2009

ABSTRACT

Initial NBTI degradation is often explained by elastic hole trapping which also considerably distorts long-term measurements. In order to clarify this issue, short-term NBT stress measurements are performed using different temperatures, stress voltages, and oxide thicknesses. The data shows a clear temperature activation and a super-linear voltage dependence, thereby effectively ruling out elastic hole tunneling. Rather, our data supports an explanation based on a thermally activated hole capture mechanism.

1. Introduction

Degradation of transistor device parameters, such as the threshold voltage V_{TH} and the mobility, already attracted the attention of the semiconductor industry many decades ago. When biasing the gate at higher temperature while keeping the rest of the transistor contacts grounded, negative bias temperature instability (NBTI) [1,2] is observed. The conventional explanation of the resulting degradation uses elastic hole trapping due to tunneling carrier exchange with the substrate (initial degradation) and the creation of interface states (long-term degradation) [3,4]. While [3,4] claim that processes in the short-time scale show a negligible temperature dependence, our latest results support a thermally activated tunneling mechanism [5].

To better understand the underlying mechanisms of short-term NBTI degradation [6] an extensive study of the short stress time behavior from the range of µs to s is necessary. Unfortunately, due to noise, accurate measurements in these time scales are difficult [7,8]. In particular, the noise in the µs regime makes it difficult to extract information on the smallest time-constants contributing to the degradation.

Currently, three fast measurement methods are used for NBTI evaluation [9]: (i) the fast-V_{TH} method [10] shortly interrupts the stress (µs delay) to quickly record V_{TH} during recovery. (ii) The fast-I_D method [11,12,6,13,14] works similarly to the fast-V_{TH} method but instead monitors the drain current I_D near V_{TH} which is then converted to ΔV_{TH} [6] using an initial $I_D(V_G)$ curve. This characteristic is only recorded around V_{TH} so as not to prestress the device. (iii) The on-the-fly (OTF) method [15,16,1] records the degradation during stress and hence does not introduce unwanted recovery, but suffers from the mobility degradation, which leads to a spurious ΔV_{TH} [17,18].

As usually implemented on a parameter analyzer OTF suffers from the problem of the initial reference measurement that already prestresses the device before the actual stress starts. In contrast, the fast-V_{TH} and the fast-I_D methods can record an unstressed reference value but suffer from the delay during measurement [11,9]. Due to its non-stop recording nature, methods (i) and (ii) [11,9] can continuously monitor recovery and, thus, allow an extrapolation back to shorter measuring delays.

Based on this experience fast rectangular gate pulses were used for short-term NBTI degradation in the range of 1 µs–1 s. Using this refined measurement procedure we collect a large dataset of stress measurements encompassing different temperatures, voltages and oxide thicknesses.

2. Samples used and stress conditions

PMOSFETS from a standard 90 nm CMOS process with plasma-nitrided oxide (around 6% of nitrogen) were used. Two thin oxide devices ($t_{ox} = 1.8\ \text{nm}, 2.2\ \text{nm}$) with geometry $W/L = 10\ \mu\text{m}/0.12\ \mu\text{m}$ and one thicker oxide device ($t_{ox} = 5\ \text{nm}$) with $W/L = 10\ \mu\text{m}/0.24\ \mu\text{m}$ were used. The devices were stressed with gate voltages $V_{G,str}$ of −1.75 V, −2.00 V, −2.25 V, and −2.50 V at temperatures of 25 °C, 75 °C, 125 °C, and 175 °C.

3. Measurement equipment and setup

State-of-the-art equipment does not meet the combined resolution and measurement speed requirements of NBTI assessment.

[*] Corresponding author. Tel.: +43 1 58801/36050; fax: +43 1 58801/36099.
E-mail address: hehenberger@iue.tuwien.ac.at (Ph. Hehenberger).

0026-2714/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.microrel.2009.06.040

Instruments either meet (and exceed) the required accuracy, but are too slow to capture the fast NBTI degradation transients (e.g. parameter analyzers), or deliver the necessary time resolution, but are limited by their inherent coarse amplitude resolution (e.g. digital storage oscilloscopes, DSO). Since in the latter case the amplitude resolution can be enhanced by averaging, while in the former there is no remedy for a too slow measurement, we use a DSO to record multiple NBTI processes and take an average of these. Care has to be taken to conform to the preconditions of proper averaging, namely to record the *same* process many times. Only in this way, the measurement noise is reduced, while the 'hidden' deterministic process is reproduced without introducing systematic errors. In our measurements this is ensured by very short stress times, and a very low duty cycle in order to achieve full relaxation in-between stresses.

The basic setup is described in [9] and uses a Hewlett-Packard 81101A pulse generator and a Tektronik TDS5034B digital storage oscilloscope. It was extended to perform short-term stress measurements including a fast gate-pulse mode and a differential amplifier.

To obtain the required resolution of better than 10^{-4} in I_D, the equipment was designed to deliver a settled gate stress voltage $V_{G,str}$ within ±1 mV in 1 μs. For this reason, a battery using a passive voltage divider and a fast electronic switch are used. As a second measure to suppress noise, the I_D of the device under test (DUT) is compared to a reference current, giving only differences, which can be captured with higher resolution prior to digitization.

According to [9] the degradation of mobility is small for stress times below 10 s. Also, in the technologies investigated, the impact of the gate current on the measurement results was found to be negligible. The recorded I_D-shift is thus regarded as due to a V_{TH}-shift alone.

4. Pulse settings

In order to automatically perform the required averaging of the recorded I_D, rectangular gate pulses were used for short-term NBTI stresses in the range of 1 μs–1 s. Each gate pulse was followed by a 100 times longer recovery sequence which allowed for full recovery of the built-up degradation [19].

Consequently, we use a pulse train with $t_{lead} = t_{trail} = 5$ ns, a width $t_W = t_{str}$ and a period of $t_P = 100\ t_{str}$, consisting of N pulses. The product Nt_P is only limited by the overall contingent measurement time $t_M = Nt_P$. A compromise between the recovery time in-between pulses ($\approx\ t_P$) to let the device fully recover and a reasonably high N has to be found in order to gain sufficient measurement accuracy through averaging.

Since the oscilloscope uses a linear time scale, but NBTI stress must be assessed on a logarithmic scale spanning at least 3 to 4 decades, we had to split the stress time of 1 s into three intervals. The according values of t_{str}, t_P, and N are shown in Table 1, as well as the resolution, which also equals the minimum stress time of the respective stress sequence.

In order to combine the three sequences into a single degradation curve with a maximum effective resolution from 1 μs to 1 s, the three stress sequences are chosen to overlap for at least one decade of time. Since only differences of currents (I_D) are recorded, the overlap regions provide information to align the sequences to a single stress characteristic. An example is displayed in Fig. 1. The offset is due to different DSO settings in each measurement sequence.

Table 1
Details of the rectangular stress pulses used to maximize the amount of recorded information together with the resolution.

Sequence	$t_W = t_{str}$	t_P	N	Resolution
1	1 ms	0.1 s	1000	0.16 μs
2	100 ms	10 s	10	16 μs
3	1000 ms	100 s	5	160 μs

5. Data extraction

Since both the measurement equipment and the pulse generator are operated at their limits, a few points have to be carefully considered during the final data extraction.

5.1. Gate voltage criteria

Monitoring V_G gives insight into the time evolution of the actual waveform, which has to be checked carefully [11]. A deep analysis of the times recorded reveals that the pulse length is around 0.3% longer than originally set by the pulse generator. This factor has to be accounted for and the real stress times t_{str} of the sequences need to be extracted using the applied gate pulse. As shown in Fig. 2 the pulse is affected by the transient behavior and a possible overshoot due to the non-instantaneous switching between $V_{G,rel}$, which is applied in-between the pulses, and $V_{G,str}$. Therefore, after the transition regime, a steady state value of $V_{G,str}$ is determined and set as $V_{G,ref}$ (usually taken at $t_{str}/2$). Then an error criterion, i.e. $|V_{G,str} - V_{G,ref}|/V_{G,ref} \leqslant \pm\epsilon$ is employed. Since noise is apparent in all three sequences, ϵ has to be chosen large enough to not disrupt the pulse, usually in the range of $\epsilon \approx 0.3\%$. Starting at $V_{G,ref}$ and moving as well to lower (to the beginning of the pulse) and higher (to the end of the pulse) times sets new borders of our accepted stress time t_{str}.

A second possibility to determine t_{str} is to skip the first datapoints during the transient until a specific time t_{skip}. This method, displayed in Fig. 2, is far easier to implement and gives stable results for various values of t_{skip}. Anyway, it suffers from the fact that for every measurement t_{skip} has to be adjusted manually. Hence, the first method is chosen.

5.2. Offset

Acquisition of 25 kSamples yields 3 to 4 usable decades in time for each sequence. The combined sequences result in 5 to 6 decades in time, with a possibly too large deviation of $V_{G,str}$ from

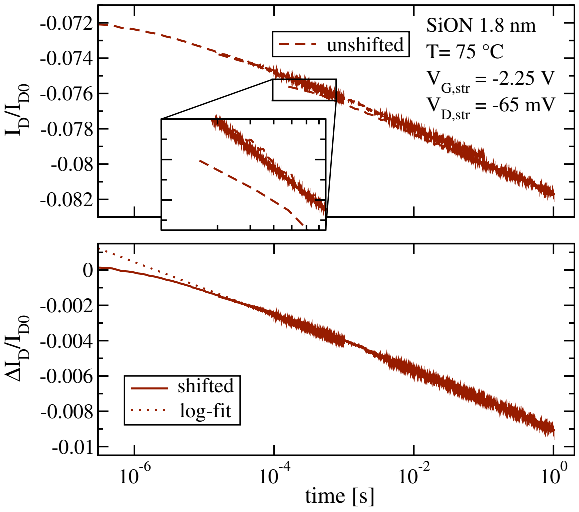

Fig. 1. Top: different DSO settings are responsible for the vertical offset. This has to be corrected to make the stress sequences coincide. Bottom: merged stress sample using a log-fit and shifted to the reference time $t_{0,ref} = 2$ μs.

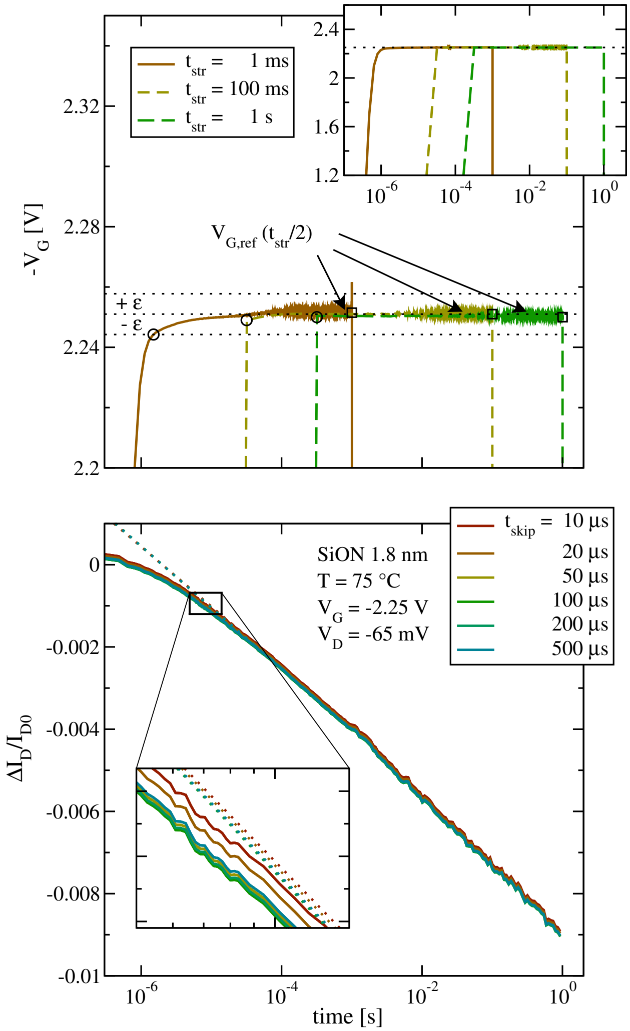

Fig. 2. Top: the inset shows the gate stress pulses. The main graph is enlarged to make the transient and the overshoot visible. This is due to the limited switching speed of the oscilloscope when moving from $V_{\mathrm{G,rel}}$ to $V_{\mathrm{G,str}}$ and back. The therefore employed error criterion ($|V_{\mathrm{G,str}} - V_{\mathrm{G,ref}}|/V_{\mathrm{G,ref}} \leqslant \pm\epsilon$) is displayed for $\epsilon = 0.3\%$. The first (last) proper values of the pulse for each sequence are marked by circles (squares). The noise is apparent in all three sequences and limits ϵ to excessively small values. Bottom: logarithmically weighting in time and skipping the first datapoints corresponding to the here varied parameter time t_{skip} does affect the shifting stability but only slightly changes the shift $\Delta I_{\mathrm{D}}/I_{\mathrm{D0}}$ (1%).

$V_{\mathrm{G,ref}}$ during the first decade. In the remaining decades the data can be either fit by a logarithmic time dependence

$$\frac{\Delta I_{\mathrm{D}}(t_{\mathrm{str}})}{I_{\mathrm{D0}}} = \frac{I_{\mathrm{D}}(t_{\mathrm{str}}) - I_{\mathrm{D0}}}{I_{\mathrm{D0}}} = -B\,\log_{10}(t_{\mathrm{str}}/t_{0,\mathrm{ref}}) \tag{1}$$

with $I_{\mathrm{D0}} = I_{\mathrm{D}}(t_{0,\mathrm{ref}})$, or a power-law $-A\,(t_{\mathrm{str}}/t_{0,\mathrm{ref}})^n$ with a very small exponent $n \approx 0.04$. I_{D0} is obtained at stress-level with a delay $t_{0,\mathrm{ref}}$ and thus *not* equal to $I_{\mathrm{D}}(0)$ [20] and results in an offset of the relative degradation, see Fig. 3.

5.3. Initial measurement as ultimate reference?

Unfortunately, the transition from the end of stress to the following recovery is always accompanied by some delay and finite transition times. Effects faster than 1 µs are not visible in our experiments. The delay of the first measurement equal to the initial measurement is broadly discussed in literature [10,20,13,21].

Some [4,8,22] argue 1 µs to be probably sufficiently short enough.

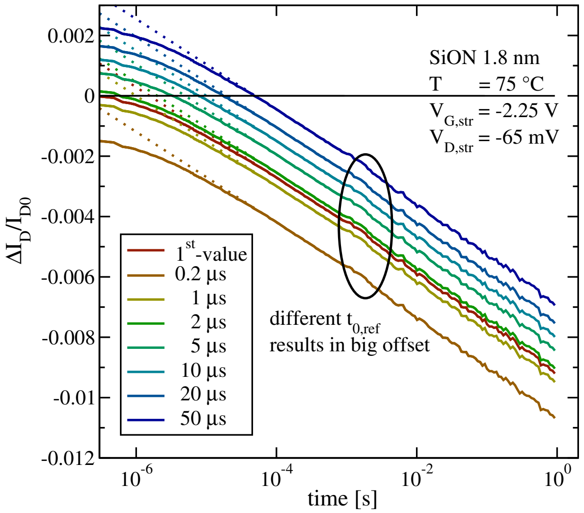

Fig. 3. Different reference times $t_{0,\mathrm{ref}}$ result in different degradation. It can be seen that for $t_{0,\mathrm{ref}} = 50$ µs about 25% of the $\Delta I_{\mathrm{D}}/I_{\mathrm{D0}}$ are missed. On the other hand, too short $t_{0,\mathrm{ref}}$ are not reasonable and result in a spurious shift by a not-yet steady measurement signal ($t_{0,\mathrm{ref}} = 0.2$ µs, 1 µs). Compare with Fig. 2.

However, as setting the reference time $t_{0,\mathrm{ref}}$ different to zero ($t_{0,\mathrm{ref}} = 0$ would be the ideal case) depends on the used equipment, different values are obtained which strongly influence the following stress behavior (Fig. 3).

5.4. Final setting of parameters

The finally extracted data is more or less sensitive to the values of the parameters $t_{0,\mathrm{ref}}$ and ϵ. For ϵ a value of 0.3% is used. As can be seen in Fig. 2 a $t_{0,\mathrm{ref}}$ slightly after the first value should be selected to both eliminate the influence of the first noisy points and delay time. Hence, $t_{0,\mathrm{ref}} = 2$ µs appears a reasonable compromise.

6. Discussion

In order to understand the microscopic physics behind the short-time degradation, the temperature, voltage, and oxide thickness dependence of the prefactor B is investigated.

6.1. Temperature scaling

The temperature dependence of $\Delta I_{\mathrm{D}}/I_{\mathrm{D0}}$ is displayed in Fig. 4 for the thinnest device ($t_{\mathrm{ox}} = 1.8$ nm) with $V_{\mathrm{G,str}} = -2.25$ V. In the range 25–125 °C, the data can be perfectly fit by a logarithmic time dependence (differences would not be visible in the plots). A slight deviation is observed for higher temperatures for $t_{\mathrm{str}} > 10$ ms, possibly due to the onset of the mechanism responsible for the long-time power-law behavior with a larger power-law exponent $n \approx 0.12$.

Apart from that, different temperatures can be scaled well to the data at $T_{\mathrm{ref}} = 175$ °C, as shown by the dotted lines in Fig. 4, and the indicated scaling factors marked by arrows.

6.2. Voltage scaling

The voltage dependence is depicted for $t_{\mathrm{ox}} = 1.8$ nm and $T = 175$ °C (Fig. 4). Scaling to $V_{\mathrm{G,ref}} = -2.50$ V leads to perfect congruence. Again, the scaling factors are shown next to their corresponding values.

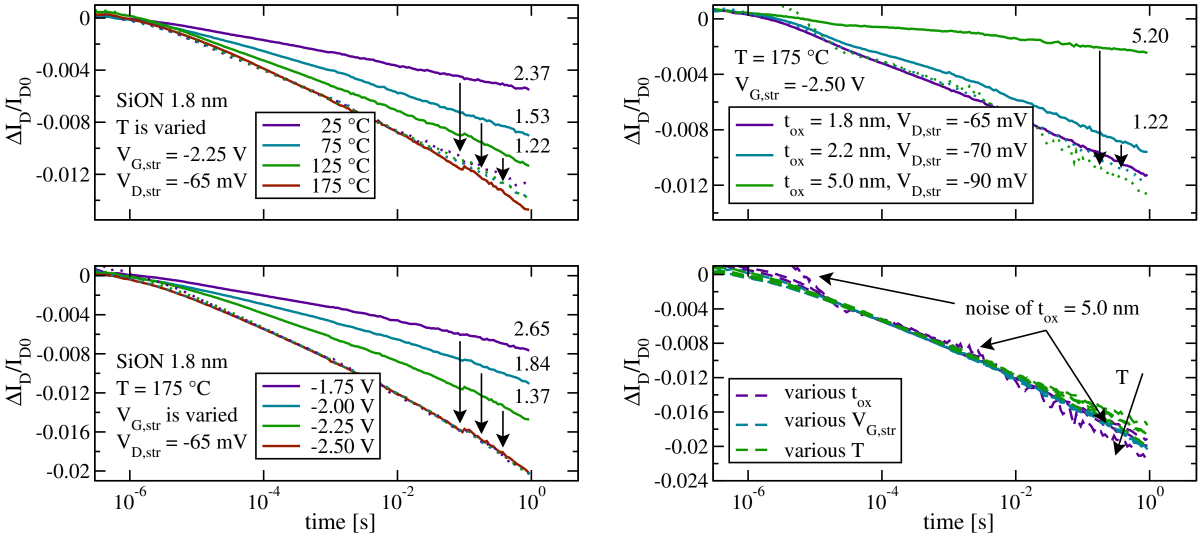

Fig. 4. Left: the temperature (25 °C, 75 °C, 125 °C, and 175 °C) and voltage dependence (−1.75 V, −2.00 V, −2.25 V, and −2.50 V) of $\Delta I_D/I_{D0}$ degradation Scaling to the dotted lines works perfectly for the later case, while different temperatures lead to a small deviation for $t_{str} > 10$ ms. The scaling factors are also given. Right: $\Delta I_D/I_{D0}$ for different oxide thicknesses (1.8 nm, 2.2 nm, and 5.0 nm) can be scaled as well. Only the thick device is affected by noise due to the low degradation. The graph at the very bottom combines the three dependencies.

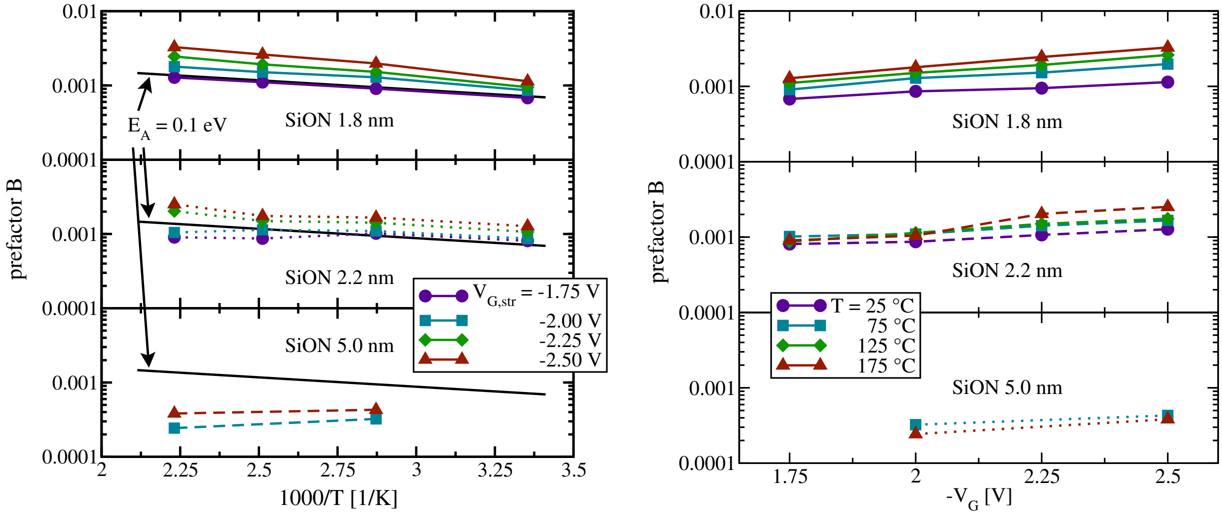

Fig. 5. Left: Arrhenius plot of the prefactor B of the log-fit, extracted from three different t_{ox} for different $V_{G,str}$. An activation energy E_A of about 0.1 eV is gained for $t_{ox} = 1.8$ nm and $t_{ox} = 2.2$ nm, represented by the black solid line. Degradation for the $t_{ox} = 5.0$ nm devices was too noisy due to too low $E_{ox} \sim (V_{G,str} - V_{TH})/t_{ox}$. Scale is equal for all plots. Right: prefactor B of the log-fit plotted for different t_{ox} with different temperature T. While $t_{ox} = 1.8$ nm shows a clear temperature activation, $t_{ox} = 5.0$ nm does not due to the low electric stress field. For $t_{ox} = 2.2$ nm the transition of the temperature dependence is visible at $T = 175$ °C between $V_{G,str} = -2.00$ V and $V_{G,str} = -2.25$ V.

6.3. Oxide thickness scaling

Due to the relatively low $\Delta I_D/I_{D0}$ degradation for $t_{ox} = 5.0$ nm resulting from the low-voltage stress conditions studied here (small E_{ox}), noise seriously limits the accuracy. Nonetheless, good scalability for different t_{ox} devices (1.8, 2.2, and 5.0 nm) can be obtained (Fig. 4).

6.4. Extracted prefactors

The prefactors B of the log-fit of various $t_{ox}, V_{G,str}$, and T are displayed in Fig. 5. In agreement with previous experiments, it is observed that low $V_{G,str}$ results in small temperature activation, while $V_{G,str}$ larger than the operating voltage gives a notable activation energy of 0.1 eV. Note that this value is in agreement with activation energies extracted at long stress times [6]. Fitting the data to a power-law $A(t_{str}/t_{0,ref})^n$ results in a exponent $n \approx 0.04$ for short-term, roughly a third of the often reported $n \approx 0.12$ of the long-term behavior.

The lower graph of Fig. 5 represents the prefactor B plotted for different t_{ox} with different temperature T. In the devices with $t_{ox} = 1.8$ nm, all the stress voltages are above the operating voltage and result in a marked temperature activation. For $t_{ox} = 2.2$ nm the transition from no temperature activation to temperature activation is observed for $T = 175$ °C between $V_{G,str} = -2.00$ V and $V_{G,str} = -2.25$ V. For the thickest oxides used in this study, $t_{ox} = 5.0$ nm, the applied stress fields are very small, resulting in no temperature activation.

All these dependencies support thermally activated tunneling mechanism [5] rather than elastic (and thus temperature-independent) hole tunneling [3].

7. Conclusions

Ultra-fast short-time NBT stress measurements from the μs to s regime using different temperatures, stress voltages, and oxide thicknesses have been performed. In this initial degradation phase, the data can be well fit by logarithmic time dependence [10,9,8]. Alternatively, a power-law using an exponent considerably smaller ($n \approx 0.04$) than generally observed during long-time stress ($n \approx 0.12$) could be used. On the other hand, the extracted activation energy of about 0.1 eV is compatible with the values typically obtained during long-time stress [6]. Finally, the extracted temperature and voltage dependencies rule out elastic and thus temperature-independent hole tunneling as being responsible for short-time NBT degradation as proposed by Denais et al. [3,4]. Another possible explanation could involve an inelastic tunneling process [5].

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme under Grant Agreement No. 216436 (project ATHENIS).

References

[1] Huard V, Denais M, Parthasarathy C. NBTI degradation: from physical mechanisms to modelling. Microelectron Reliab 2006;46(1):1–23.
[2] Schroder D, Babcock J. Negative bias temperature instability: road to cross in deep submicron silicon semiconductor manufacturing. J Appl Phys 2003;94(1):1–18.
[3] Denais M, Huard V, Parthasarathy C, Ribes G, Perrier F, Revil N, et al. Interface trap generation and hole trapping under NBTI and PBTI in advanced CMOS technology with a 2-nm gate oxide. T-DMR 2004;4:715–22.
[4] Mahapatra S, Maheta V, Islam A, Alam M. Isolation of NBTI stress generated interface trap and hole-trapping components in PNO p-MOSFETs. T-ED 2009;56(2):236–42.
[5] Grasser T, Kaczer B, Gös W, Aichinger Th, Hehenberger Ph, Nelhiebel M. A two-stage model for negative bias temperature instability. IRPS 2009:33–44.
[6] Grasser T, Kaczer B. Evidence that two tightly coupled mechanisms are responsible for negative bias temperature instability in oxynitride MOSFETs. T-ED 2009;56(5):1056–62.
[7] Maheta V, Kumar E, Purawat S, Olsen C, Ahmed K, Mahapatra S. Development of an ultrafast on-the-fly I_{DLIN} technique to study NBTI in plasma and thermal oxynitride p-MOSFETs. T-ED 2008;55(10):2614–22.
[8] Zhang J, Ji Z, Chang M, Kaczer B, Groeseneken G. Real Vth instability of pMOSFETs under practical operation conditions. IEDM 2007:817–20.
[9] Reisinger H, Brunner U, Heinrigs W, Gustin W, Schlünder C. A comparison of fast methods for measuring NBTI degradation. T-DMR 2007;7(4):531–9.
[10] Reisinger H, Blank O, Heinrigs W, Gustin W, Schlünder C. A comparison of very fast to very slow components in degradation and recovery due to NBTI and bulk hole trapping to existing physical models. T-DMR 2007;7(1):119–29.
[11] Kaczer B, Grasser T, Roussel PhJ, Martin Martinez J, O'Connor R, O'Sullivan B, et al. Ubiquitous relaxation in BTI stressing new evaluation and insights. IRPS 2008:20–7.
[12] Kaczer B, Arkhipov V, Degraeve R, Collaert N, Groeseneken G, Goodwin M. Disorder-controlled-kinetics model for negative bias temperature instability and its experimental verification. IRPS 2005:381–7.
[13] Grasser T, Kaczer B, Hehenberger P, Gös W, O'Connor R, Reisinger H, et al. Simultaneous extraction of recoverable and permanent components contributing to bias-temperature instability. IEDM 2007:801–4.
[14] Shen C, Li M-F, Wang X, Yeo Y-C, Kwong D-L. A fast measurement technique of MOSFET Id-Vg characteristics. EDL 2006;27(1):55–7.
[15] Krishnan A, Reddy V, Chakravarthi S, Rodriguez J, John S, Krishnan S, et al. NBTI impact on transistor & circuit: models, mechanisms & scaling effects. IEDM 2003:1–4.
[16] Denais M, Bravaix A, Huard V, Parthasarathy C, Ribes G, Perrier F, et al. On-the-fly characterization of NBTI in ultra-thin gate oxide PMOSFETs. IEDM 2004.
[17] Grasser T, Wagner P-J, Hehenberger Ph, Gös W, Kaczer B. A rigorous study of measurement techniques for negative bias temperature instability. T-DMR 2008;8(3):526–35.
[18] Islam A, Kumar E, Das H, Purawat S, Maheta V, Aono H, et al. Theory and practice of on-the-fly and ultra-fast VT measurements for NBTI degradation: challenges and opportunities. IEDM 2007:805–8.
[19] Rangan S, Mielke N, Yeh E. Universal recovery behavior of negative bias temperature instability. IEDM 2003:341–4.
[20] Shen C, Li M-F, Foo C, Yang T, Huang D, Yap A, et al. Characterization and physical origin of fast Vth transient in NBTI of pMOSFETs with SiON dielectric. IEDM 2006.
[21] Schlünder C, Vollertsen R-P, Gustin W, Reisinger H. A reliable and accurate approach to assess NBTI behavior of state-of-the-art pMOSFETs with fast-WLR. ESSDERC 2007:131–4.
[22] Kumar E, Maheta V, Purawat S, Islam A, Olsen C, Ahmed K, et al. Material dependence of NBTI physical mechanism in silicon oxynitride (SiON) pMOSFETs: a comprehensive study by ultra-fast on-the-fly (UF-OTF) I_{DLIN} technique. IEDM 2007:809–12.