Defect Creation Stimulated by Thermally Activated Hole Trapping as the Driving Force Behind Negative Bias Temperature Instability in SiO_2, SiON, and High-k Gate Stacks

Tibor Grasser*, Ben Kaczer°, Thomas Aichinger†, Wolfgang Goes*, and Michael Nelhiebel‡

* Christian Doppler Laboratory for TCAD at the Institute for Microelectronics, TU Wien, Austria
° IMEC, Leuven, Belgium † KAI, Villach, Austria ‡ Infineon Technologies, Villach, Austria

***Abstract*— Recent publications on negative bias temperature instability have clearly demonstrated the existence of two components contributing to the phenomenon, with one of them recovering over many timescales and the other being more or less permanent. Interestingly, these two components seem to be coupled since their effect cannot be separated by the application of different stress voltages and stress temperatures. Based on this scalability we suggest a new model which can explain the experimental data during both stress and recovery for our pure SiO_2, oxynitride, and high-k devices under a wide range of experimental conditions.**

I. Introduction

Many recent publications dealing with negative bias temperature instability (NBTI) have discussed the existence of a recoverable component observed on top of a slowly recovering or even permanent component [1–3]. Conventionally, the recoverable component is attributed to hole trapping while the permanent component is explained by the creation of interface states [1]. We have recently pointed out a *serious problem* with the conventional interpretation that two *independent* components result in the overall degradation observed during NBT stress [4, 5]. This is because these two components should have a different voltage and temperature acceleration, allowing for their separation by the application of a suitably chosen combination of stress temperatures and voltages. Quite to the contrary, however, we observed that NBTI data obtained at different stress and relaxation times, in a range of stress voltages ($-1.0\,\mathrm{V}$ ... $-2.0\,\mathrm{V}$) and stress temperatures ($25\,^\circ\mathrm{C}$... $200\,^\circ\mathrm{C}$), can often be made to overlap for over seven decades in time by multiplication with a suitably chosen scaling factor (Figs. 1 and 2).

This scalability *cannot be explained by any available model* [4]. *We show that the only explanation for this wide scalability is coupling of the above two widely-accepted mechanisms*. We further devise *the first microscopic model* that can *explain the measurement data in the full investigated voltage and temperature ranges for three vastly different technologies* (ultra-thick SiO_2, ultra-thin SiON, and high-k).

II. Scalability of NBTI data

In order to further challenge the observed scalability, we investigated two extreme technologies, one with an ultra-thick $30\,\mathrm{nm}$ SiO_2 layer and the other with a SiO_2/HfO_2 (HK) gate stack. Amazingly, the same scalability is observed and all devices show qualitatively the same stress and recovery behavior, quite contrary to the notion that NBTI in devices with ultra-thick oxides is understood and complications arise only due to the introduction of nitrogen or alternative gate stacks. This also supports the view that NBTI is primarily determined by the Si/SiO_2 interface properties rather than the oxide bulk [8].

In order to study the scalability in greater detail, the field and temperature dependences of the scaling factors have been extracted. Clearly similar scaling factors for the three technologies emerge (Fig. 3). While the field-dependence of the scaling factors has been found similar for the SiON and HK devices, their slope is somewhat larger for the ultra-thick SiO_2 device. Most remarkably, the *temperature activation is the same in all investigated technologies*.

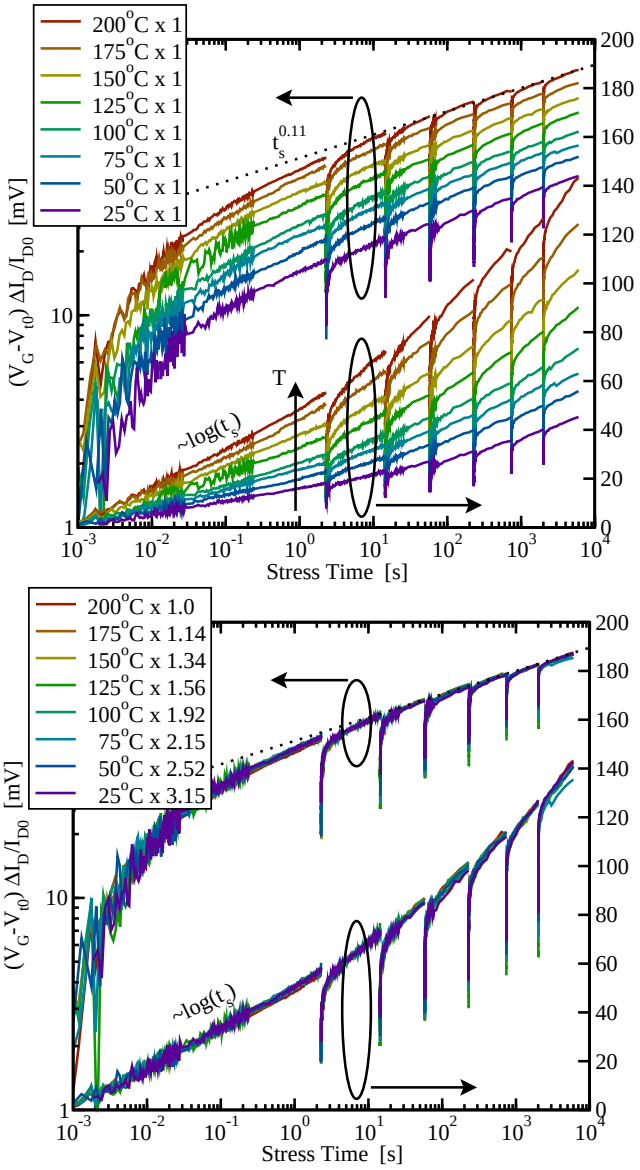

Fig. 1. Degradation of the drain current collected during subsequent stress/recovery sequences (extended MSM technique (eMSM), [3, 6] using 7 alternating stress and recovery phases). The stress temperature was varied at a fixed stress voltage of $-2\,\mathrm{V}$. The drain current is converted to $\Delta V_{\mathrm{th}}^{\mathrm{OTF}}$ using the OTF prescription (e.g. [7]). **Top**: The unscaled data initially follows a logarithmic time dependence while the long-time data may be approximated by a power-law with the exponent $n = 0.11$. The slope of the initial log shows clear temperature activation while the power-law exponent is roughly temperature-independent. **Bottom**: Multiplication of each data set with a constant value results in a *nearly perfect overlap*. The scaling factors are independent of stress time, indicating that the initial log and the long-time power-law behavior are due to a related process.

978-1-4244-2195-4/08/$25.00 ©2008 IEEE

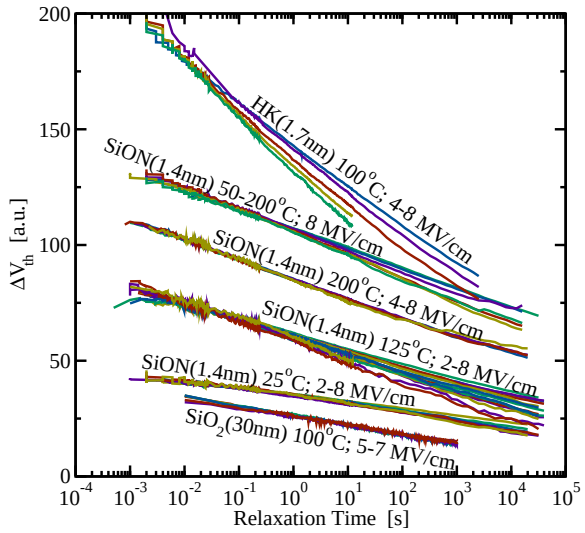

Fig. 2. Recovery of the threshold voltage shift at different temperatures, voltages, and stress times, for three completely different gate stacks, ultra-thick SiO_2, ultra-thin SiON, and HK. Amazingly, a similar scalability as during stress is observed during recovery for all technologies. All data sets show good scalability and thus do not support the idea that NBTI is a consequence of two independent mechanisms.

III. Need for very wide distributions

The *very broad scalability* implies that *NBTI is either due to a single mechanism* (which then would have to be able to explain both the recoverable and the permanent contributions) or due to *two tightly coupled mechanisms*. In our previous attempt [5] we constructed such a single-mechanism model based on the transition of hydrogen away from the dangling bond via an intermediate state to the transport states. This triple-well model was found to reproduce complicated stress/relaxation sequences with unprecedented accuracy.

However, when applied with a fixed configuration of barriers to a larger set of stress/relaxation sequences obtained at different T and V_{stress}, a very large variance of the Gaussian-distributed barriers in the model would be required in order to reproduce the ubiquitous log-like recovery behavior of ΔV_{th} (i.e., $\partial \Delta V_{\mathrm{th}}/\partial t \sim 1/t$) [3]. Such a large variance, which is basically equivalent to a uniform distribution in energy, is unusual for a transition of hydrogen to an intermediate state, but has been applied since the 1970s in attempts to understand hole trapping, $1/f$-noise, and thermally stimulated currents at semiconductor surfaces [10, 14]. In these models it is assumed that holes can be captured via a *thermally activated* trapping process in near-interfacial states, for instance into oxygen vacancies (Si–Si bonds in SiO_2). The thermal activation is related to the variation in the bond-length of the strained Si–Si bonds in the amorphous interfacial layer. This strongly suggests that the *triple-well should be split into two coupled wells*, thereby honoring the different physical origins of the recoverable and permanent components.

IV. New Model

Based on the above observations we formulate a *new model for NBTI* (see Fig. 4): *First*, upon application of stress, *holes can be trapped* into near-interfacial states via a thermally activated process. As will be shown later, such a thermally activated hole capture can naturally explain many features observed during NBTI, most notably the temperature-dependent initial log-like degradation and recovery (consequences of the near-uniform barrier-height distribution). The near-interfacial states are assumed to exist in two flavors, once as mere hole traps and once as precursors for defect creation. Potential

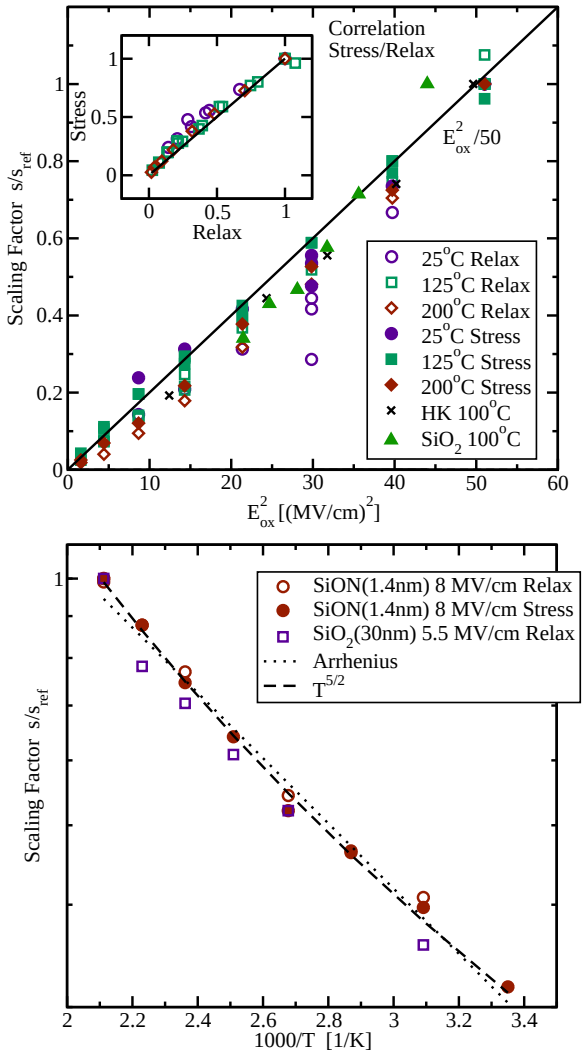

Fig. 3. **Top**: The scaling factors obtained for the three different technologies as a function of the stress field can be described by an exponential dependence [9]. The field dependence of all technologies is the same and follows E_{ox}^2. The inset shows the excellent correlation between the scaling factors obtained during stress and recovery for the SiON device. **Bottom**: The temperature dependence of the scaling factors was found to be *the same for all technologies* and can be approximated by an Arrhenius law. A power-law fit $T^{5/2}$ approximating the numerical solution of the model suggested here is in very good agreement with the data and close to the analytical approximation of T^2 given in (8).

candidates for such states would be the various oxygen vacancies [14] (E'-centers) and the silicon-hydrogen bridge [14] Si–H–Si, respectively. *Second, once a hole is trapped in the defect precursors*, the *release of* H *is considerably enhanced* due to the weakening of the binding energies, which eventually results in the creation of poorly recoverable defects, e.g. P_{b}-centers in SiO_2 layers and K_{N}-centers in oxynitrides [15]. The important aspect of this model is that *hole capture is not assumed to happen instantaneously as e.g. in [13], but rather via a thermally activated, dispersive process triggering defect creation*. As a consequence, the created interface states ΔN_{it} become coupled to the charges created by hole trapping, ΔN_{ot}. Mathematically, the above can be expressed by two coupled double-wells with occupancies f_1 and f_2

$$\dot{f}_1 = -f_1 k_{1\mathrm{f}} + (1-f_1) k_{1\mathrm{r}} \quad (1)$$

$$\dot{f}_2 = -(1-f_1) f_2 k_{2\mathrm{f}} + (1-f_2) k_{2\mathrm{r}} \quad (2)$$

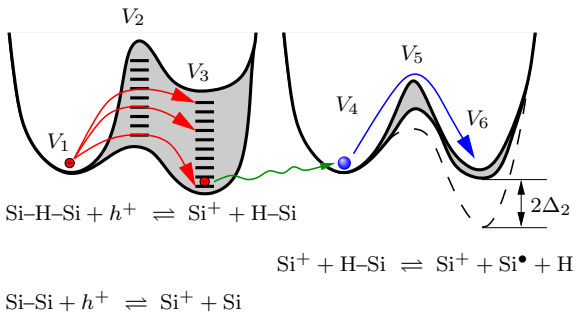

Fig. 4. *The new model* suggested for the explanation of the experimental data: Initially, charges are captured (for instance into a Si–H–Si bridge) via a thermally activated process (represented by double-well $V_1 - V_2 - V_3$). A captured hole accelerates the release of the hydrogen atom which then creates a dangling bond (represented by double-well $V_4 - V_5 - V_6$). The hole trapping sites are assumed to be uniformly distributed in a wide energy range [10] (widths $2\sigma_2$ and $2\sigma_3$), while the energy V_5 required for the release of a hydrogen atom follows a much narrower Gaussian distribution ($\sigma_5 \approx 0.1\,\mathrm{eV}$, see [1]). The hole capture forward rate is assumed to be proportional to the surface hole concentration, approximated by the temperature-independent term $\exp(-\Delta_1(E_{\mathrm{ox}}))$, while creation of interface states is accelerated by energy-level shifts proportional to the electric field [11].

where the reaction rates are given by

$$k_{1\mathrm{f}} = \nu_1 \mathrm{e}^{-\beta(V_2-V_1)}\mathrm{e}^{\Delta_1}, \quad k_{2\mathrm{f}} = \nu_2 \mathrm{e}^{-\beta(V_5-V_4-\Delta_2)} \tag{3}$$

$$k_{1\mathrm{r}} = \nu_1 \mathrm{e}^{-\beta(V_2-V_3)}, \quad k_{2\mathrm{r}} = \nu_2 \mathrm{e}^{-\beta(V_5-V_6+\Delta_2)}\,. \tag{4}$$

The observed charges are then given by

$$N_{\mathrm{ot}} = \langle 1 - f_1 \rangle\, N_{\mathrm{ot}}^0, \quad N_{\mathrm{it}} = \langle 1 - f_2 \rangle\, N_{\mathrm{it}}^0 \tag{5}$$

An approximate analytical solution for the initial behavior ($\Delta N_{\mathrm{it}} \ll \Delta N_{\mathrm{ot}}$) can be obtained as

$$\Delta V_{\mathrm{th}}^{\mathrm{OTF}}(t_{\mathrm{s}}) \approx A \log(t_{\mathrm{s}}/t_0) \tag{6}$$

$$\Delta V_{\mathrm{th}}(t_{\mathrm{s}}, t_{\mathrm{r}}) \approx A \log(1 + t_{\mathrm{s}}/t_{\mathrm{r}}) \tag{7}$$

$$A = \frac{1}{4\sigma_2\sigma_3}\frac{\mathrm{q}^2}{{\mathrm{k_B}}^2}T^2\Delta_1 \tag{8}$$

For small stresses, the analytic solution follows a logarithmic behavior during both OTF stress (with t_0 as first measurement point) and recovery, as observed in the experiments. The prefactor A depends linearly on the stress-parameter Δ_1 and quadratically on temperature (cf. Fig. 3).

Our new model needs to be compared to the model suggested by Huard *et al.* [1] which can produce excellent fits for a single temperature and stress-voltage [1, 4]. However, Huard's model uses the hole-trapping model of Tewksbury [16], which is based on elastic tunneling into traps located at various distances away from the interface and is consequently nearly *temperature-independent*, in contradiction to our data. Furthermore, it cannot explain long recovery tails in the order of 10^5 s in ultra-thin oxides with a physical meaningful set of parameters [4], as there hole detrapping would occur at timescales in the order of milliseconds. Finally, mere superposition of such a hole trapping mechanisms with the strongly temperature-dependent defect creation process *contradicts the observed very broad scalability*.

V. Results and Discussion

In order to evaluate the new model against the three technologies, their average degradation behavior had to be extracted. This has been done by using devices of each technology stressed at different temperatures and voltages. For each technology, the samples where calibrated together with field-dependent values of the stress parameters Δ_1 and Δ_2, while the temperature dependence is implicitly formulated in the model (we remark that a different Ansatz may result in temperature-dependent Δ_i). Thus, for a correct extraction of the barriers, the samples had to contain devices stressed with the same field at different temperatures. In order to keep the extraction simple, all devices of each technology where assumed to have the same number of oxide trap precursors N_{ot}^0 and the same number of defect precursors N_{it}^0. In fact, this seems to be accurate within 10%. Consequently, a comparison of the measurement data with the simulation is expected to give the average response to the stressing conditions and may slightly deviate for a particular measurement.

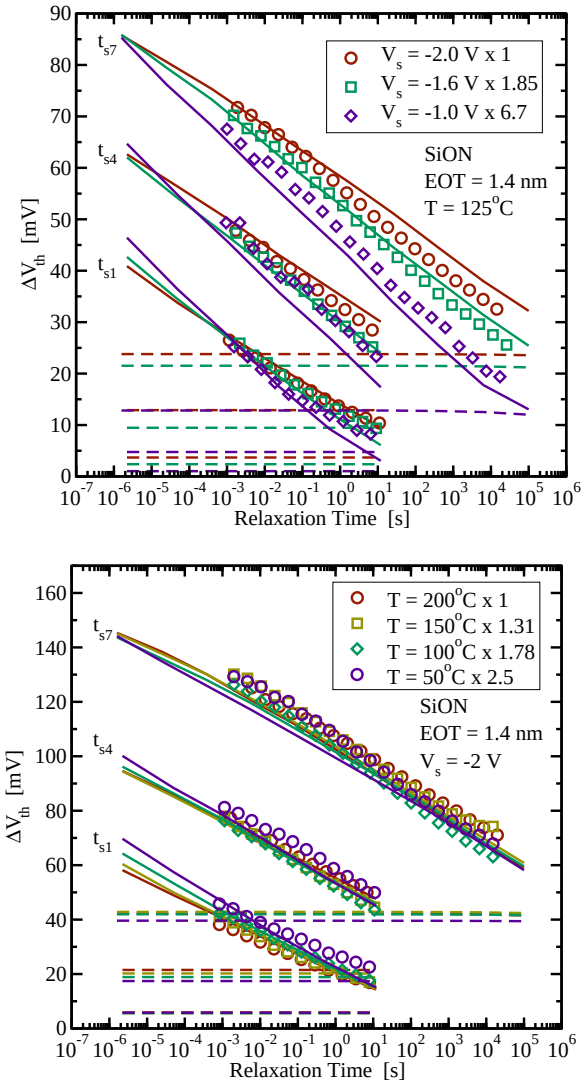

Fig. 5. Verification of the scalability of the model. **Top**: Voltage scaling: the degradation resulting from three different bias conditions is compared to the measurement data. For better visibility, the data is normalized to the simulated recovery after the last stress phase at $t_{\mathrm{r}} = 2\,\mu$s. Clearly, the decrease in slope for increasing stress is well reproduced. Keep in mind that the barriers are determined for the average case and each individual measurement may slightly deviate from this nominal behavior. **Bottom**: Temperature scaling: The model also scales properly at different temperatures.

The simulated SiON devices are compared to measurements in Figs. 5–6. *Unprecedented accuracy is obtained during both stress and relaxation* for each individual device at all temperatures and voltages, the *scaling property is excellently reproduced*, and also the *acceleration of recovery as a response to positive bias* can be well described. Results of similar accuracy were obtained for the HK device (see Fig. 7) and the ultra-thick devices Fig. 8. Finally, the

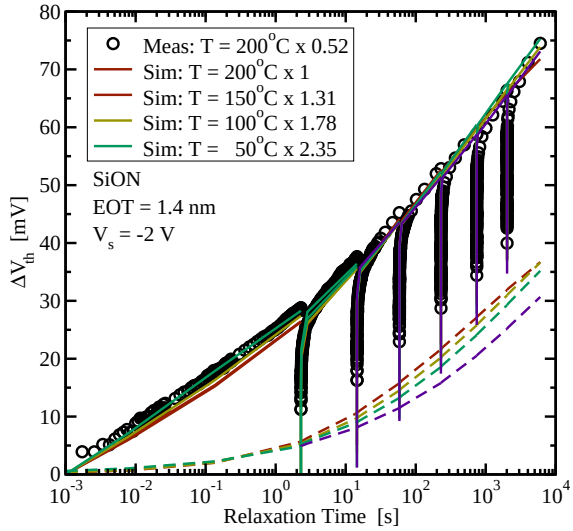

Fig. 6. Comparison of the model with the measurement data obtained during the stress phases (i.e. "OTF") of the eMSM sequence (cf. Fig. 1). While the model scales perfectly, the estimated $\Delta V_{\rm th}$ from the OTF measurement had to be scaled by 0.52. This is ascribed to the incomplete conversion of OTF $\Delta I_{\rm D}$ to $\Delta V_{\rm th}$ which neglects the mobility variations [12, 13].

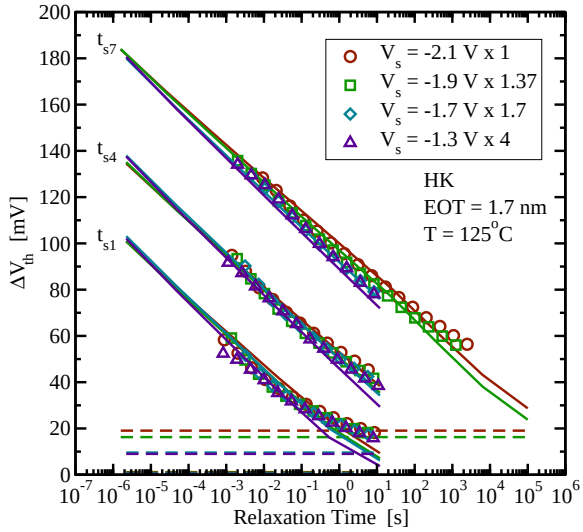

Fig. 7. Comparison of the model calibrated to 9 high-k devices. Shown is the scaling of the recovery data after three selected stress times obtained at different stress voltages. The scaling property is excellently reproduced by the model.

field-dependence of the extracted barrier for the creation of interface states is shown in Fig. 9, where all technologies show comparable behavior.

VI. Conclusions

We have suggested *a new dispersive model for negative bias temperature instability* based on *defect creation stimulated by thermally activated hole trapping.* This model can explain degradation and recovery over a wide range of bias voltages sequences and stress temperatures. Excellent agreement with measurements from three vastly different technologies (SiO_2, SiON, and HK) is obtained, underlying that NBTI is driven by a technology-independent mechanism primarily related to the chemistry of the silicon/silicon-dioxide interface and near-interfacial layer.

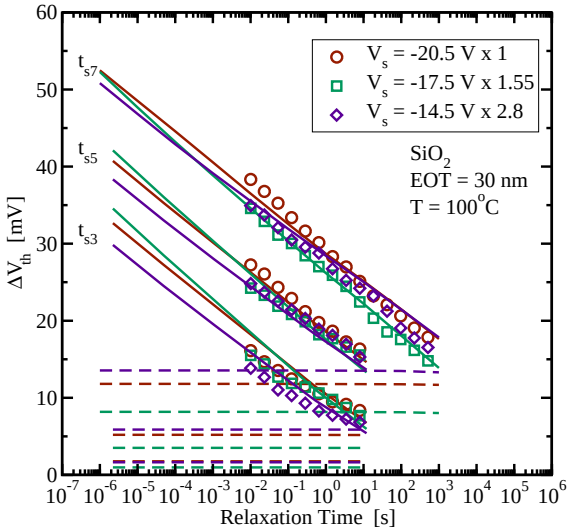

Fig. 8. Scaled relaxation data of the SiO_2 devices after stresses at different voltages at the same temperature compared to the simulation. As for the SiON samples, excellent scalability is obtained for the SiO_2 devices.

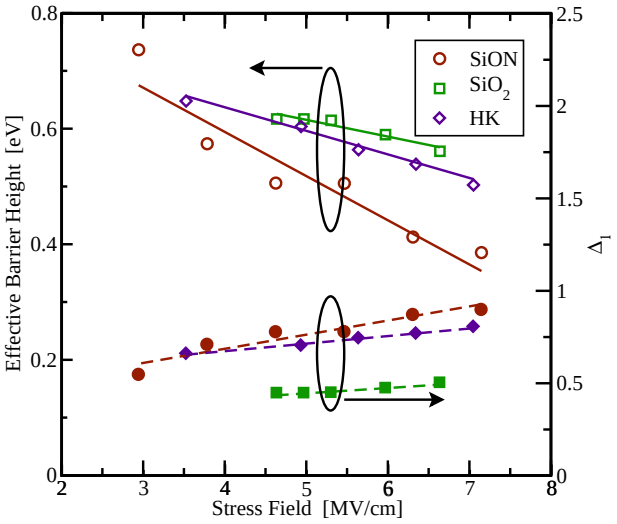

Fig. 9. The extracted field acceleration parameters for the three different technologies. Shown is the mean of the field-dependence of the first barrier of the permanent component, $V_5 - \Delta_2(E_{\rm ox})$, and the parameter Δ_1. All three technologies are comparable.

References

[1] V. Huard *et al.*, in *IEDM* (2007), pp. 797–800.
[2] C. Shen *et al.*, in *IEDM* (2006), pp. 333–336.
[3] B. Kaczer *et al.*, in *IRPS* (2008), pp. 20–27.
[4] T. Grasser, in *IRPS* (2008), (Tutorial).
[5] T. Grasser *et al.*, in *IRPS* (2008), pp. 28–38.
[6] B. Kaczer *et al.*, in *IRPS* (2005), pp. 381–387.
[7] S. Mahapatra *et al.*, in *IRPS* (2007), pp. 1–9.
[8] S. Zafar *et al.*, in *VLSI Symp.* (2006), pp. 23–25.
[9] A. Islam *et al.*, T-ED **54**, 2143 (2007).
[10] M. Weissman, Rev.Mod.Phys **60**, 537 (1988).
[11] J. McPherson, in *IRPS* (2007), pp. 209–216.
[12] T. Grasser *et al.*, in *IIRW* (2007), pp. 6–11.
[13] A. Islam *et al.*, APL **90**, 083505 (2007).
[14] D. Fleetwood *et al.*, T-ED **49**, 2674 (2002).
[15] J. Campbell *et al.*, T-DMR **7**, 540 (2007).
[16] T. Tewksbury *et al.*, JSSC **29**, 239 (1994).

Questions and Answers

Q: (Slide 9) For the hole trap, have many different excited states. If the hole is trapped in excited state, is the hole in a localized state (bound state)?
A: The detailed physics of this hole capture process and whether the behavior observed during NBTI is compatible with the considerable amount of literature available on the E' centers is still open. A more detailed description, where these essential details will have to be considered, is currently under development.