Microelectronics Reliability 52 (2012) 1883–1890

Contents lists available at SciVerse ScienceDirect

Microelectronics Reliability

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

Defect-centric perspective of time-dependent BTI variability

M. Toledano-Luque [a,*], B. Kaczer [a], J. Franco [a,b], Ph.J. Roussel [a], T. Grasser [c], G. Groeseneken [a,b]

[a] Imec, Kapeldreef 75, B-3001 Leuven, Belgium
[b] ESAT, KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[c] Institute for Microelectronics, Technische Universität Wien, Gußhausstraße 27-29, A-1040 Vienna, Austria

ARTICLE INFO

Article history:
Received 30 May 2012
Accepted 27 June 2012
Available online 25 July 2012

ABSTRACT

With the continuous downscaling of CMOS device dimensions, (i) The number of gate oxide defects in each device decreases to a numerable level, while their relative impact on the device characteristics increases. (ii) The properties of each defect, such as its capture and emission times and its impact, are voltage and/or temperature dependent and widely distributed. (iii) The occupation kinetics of each defect is known to be stochastic. All of these result in each of the nominally identical nm-scaled devices behaving very differently during operation, resulting in increasing time-dependent variability (heteroskedasticity). Consequently, the lifetime of nm-sized devices cannot be predicted individually, but can be described in terms of time- (or workload-) dependent distributions.

1. Introduction

As the vertical scaling of metal–oxide–semiconductor field effect transistor (MOSFET) devices continues, the oxide electric field increases and the so-called Bias Temperature Instability (BTI) becomes one of the most critical factors, complicating the qualification of the future technology nodes [1–3]. Furthermore, the number of stochastically behaving gate oxide defects in each device decreases to a numerable level due to the lateral downscaling, while their relative impact on the device characteristics increases. For all these reasons, BTI lifetime cannot be described any longer by a unique number, and *BTI lifetime distribution* has to be taken into consideration. As a consequence, even in the ideal case of the *average BTI lifetime* meeting the ITRS [4] specifications, a fraction of nanoscaled devices will fail at low overdrives. In this paper, the necessary physical understanding to predict the BTI lifetime distributions is developed and is introduced into an "atomistic" circuit simulator that takes realistic workloads into account.

We start by briefly reviewing the elementary definitions and experimental observations of BTI in large area and in nm-scaled devices. Contrary to the continuous relaxation curves observed on large area devices after bias temperature stress, giant discrete threshold voltage V_{TH} shifts are measured on nanoscaled devices and linked to drain random telegraphic noise (I_D-RTN) [5]. We then demonstrate that many properties of gate oxide defects, such as characteristic emission and capture times and V_{TH} impact can be directly extracted from BTI relaxation measurements in deeply scaled devices [6]. Afterward we show how the understanding of gate oxide defect properties can be used to explain time dependent BTI variability in deeply scaled technologies. Finally, an atomistic simulator based on existing industry-standard tools is presented for circuit assessments.

2. Bias Temperature Instability BTI in large and in nm-scaled devices

During CMOS circuit operation the devices typically undergo electrical stress at elevated temperature resulting in a shift of the device parameters such as its threshold voltage, channel mobility, transconductance, and subthreshold slope, instigating a decrease of the FET's drive current. Since these *instabilities* are strongly accelerated by *temperature T* and *gate bias V_G, they are known by the acronym BTI* (*Bias Temperature Instability*). These phenomena are mainly the consequence of charging of defects in the gate oxide and at its interface [7]. BTI in *n*-channel FET devices, which are typically biased at positive V_G in CMOS circuits, is referred to as positive BTI (PBTI), while negative BTI (NBTI) takes place in *p*-channel FETs.

Fig. 1a illustrates the typical gradual shift of pFET threshold voltage ΔV_{TH} during accelerated stress at elevated temperature [8]. Data are typically measured at several V_G's to obtain the maximum circuit operating voltage V_{DD} that the devices could withstand for 10 years while the ΔV_{TH} is below a given value (typically 30 or 50 mV).

However, this extrapolation procedure is problematic due to the immediate ΔV_{TH} recovery after the stress bias is removed [5,7], as illustrated in Fig. 1b. As we will discuss henceforth, this recovery or relaxation typically proceeds on many time scales, causing difficulties to extrapolate to both shorter and longer relaxation times, and

* Corresponding author. Tel.: +32 16 281647.
E-mail address: toleda@imec.be (M. Toledano-Luque).

0026-2714/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.microrel.2012.06.120

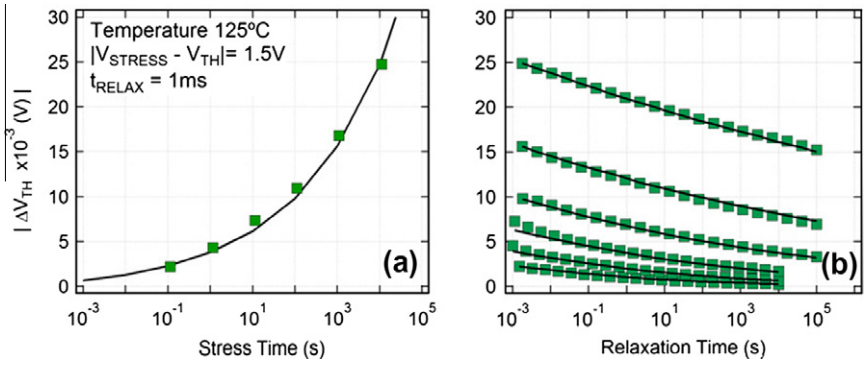

Fig. 1. (a) Threshold voltage shift ΔV_{TH} is observed during negative gate bias stress and high temperature (125 °C) in a $W \times L = 10 \times 0.5\ \mu m^2$ pFET formed by 0.8 nm-SiO_2/1.8 nm-HfSiO. (b) When the stress bias is removed a recovery of the effect is observed.

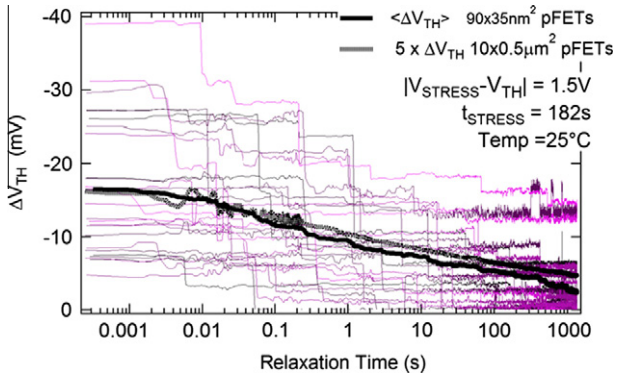

Fig. 2. Bias Temperature Instability (BTI) relaxation transients obtained on $W \times L = 90 \times 35\ nm^2$ 0.8 nm-SiO_2/1.8 nm-HfSiO pFETs. Steps due to single-carrier discharge events are evident. The large dispersion is due to the stochastic distributions of N_T and the impact of each trap. Note that the average relaxation resembles the curve taken on a large area device.

therefore to obtain the permanent degradation component [7,9,10]. This ΔV_{TH} relaxation is thus a crucial problem for BTI measurements, interpretation, and extrapolation. Understanding the recoverable component has been crucial to unraveling the BTI mechanism and has been achieved by means of the thorough study of deeply scaled devices.

Fig. 2 displays the relaxation traces after a BTI stress obtained on $90 \times 35\ nm^2$ pFETs. Each trace reveals the combined response of multiple defects and every discrete drop is due to a single-carrier discharge event [5,7]. The average relaxation resembles the curve taken on a large area device under equal stress condition, indicating that *identically behaving traps are responsible for BTI in both small and large area FETs* [8,11].

The figure illustrates the wide variation in the behavior of individual devices. We will show hereafter that this variation can be described analytically [12] by means of two parameters: the mean *total* ΔV_{TH} and the mean impact on V_{TH} per trap η, i.e., $\langle total\ \Delta V_{TH} \rangle = \eta \times N_T$, with N_T the mean number of active traps. The ΔV_{TH}'s obtained from both nanoscaled and large pFETs as a function of temperature (not shown) follow an Arrhenius law with the same activation energy [8]. However, a larger degradation was observed in small devices at all stress conditions due in part the larger impact per charged trap caused by channel percolation effects in small devices [13–15] as explained in the next sections.

3. BTI: a non-steady state case of random telegraph noise

From the quantized recovery behavior observed in nanoscaled FETs it is straightforward to understand the recoverable component of BTI as the dynamic non-steady state case of random telegraph noise (RTN) [16]. As in the case of I_D-RTN, the large quantized ΔV_{TH}'s observed in Fig. 2 are explained by the non-uniform potential at the Si/SiO_2 interface caused by the random distributions of dopants in the channel and charged traps in the dielectric. The potential fluctuations produce variations of the inversion charge density and, consequently, preferential conduction paths from the source to the drain. The charging and discharging of single oxide traps over critical positions of the conduction paths can produce significant fluctuations of the drain current [12,14]. The change of drain current can be in turn transformed into a V_{TH} shift when taking the I_D–V_G curve of the fresh device as a reference [17].

In the case of RTN, the emission and capture times are in the same order of magnitude causing random switching of the drain current at fixed V_G. In the case of BTI, the capture of charge is forced at high gate voltage (V_{STRESS}) and the emission at low voltage (V_{RELAX}). This allows studying states with dissimilar emission and capture times, reducing the prohibitive acquisition time of standard RTN experiments.

In this paper, we present two approaches for the study of the discretized relaxation curves obtained on nm-scaled devices: (i) repeatedly performing the same experiment on a single device or (ii) conducting one single experiment on many devices.

From the former approach, a new technique named *time dependent defect spectroscopy* (TDDS) [6] has been developed allowing the study of the kinetic properties of single defects as a function of stress/recovery bias conditions and temperature [4,18–21]. These studies have revealed interesting facts about the charge trapping component that we summarize in the next section.

From the latter approach, we demonstrate the methodology to predict the ΔV_{TH} *distributions* after BTI stress through a detailed understanding of the *atomistic* impact of *individual* traps [11,15,22]. This approach has proven to be useful for reliability engineers [21] and circuit designers to predict time dependent BTI variability [23].

Before studying in depth the ΔV_{TH} recovery traces, it is worth mentioning the recently found correlation (see Fig. 3) between the discrete gate (I_G-RTN) and drain current (I_D-RTN) fluctuations in nanoscaled SiON pFETs and nFETs [24,25]. This demonstrates that both effects are due to charging and discharging of the same isolated defects and, therefore, the conclusions exposed in the next sections are also applicable to defects causing I_G-RTN.

4. Kinetic of individual traps

As previously stated, a new methodology has been introduced to study the statistical properties of individual traps called time dependent defect spectroscopy (TDDS) [6]. In this section, we explain this methodology and the way to obtain the response of

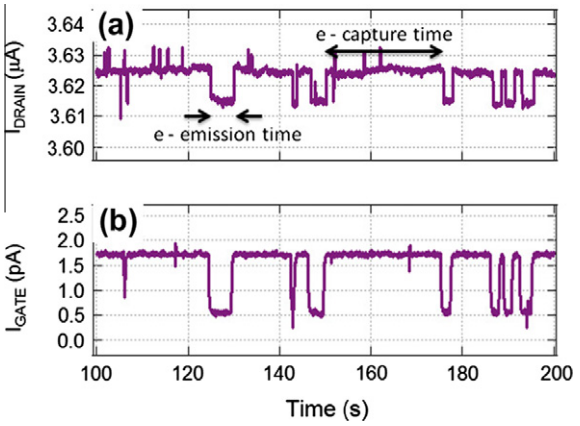

Fig. 3. (a) I_D and (b) I_G traces simultaneously registered by means of Keithley 2636 SMUs on a nanoscaled 90 × 35 nm^2 2.3 nm-EOT SiON nFET showing synchronized fluctuations and demonstrating the correlation between I_D and I_G RTN. The lower level of the drain current corresponds to the periods when the trap is negatively charged.

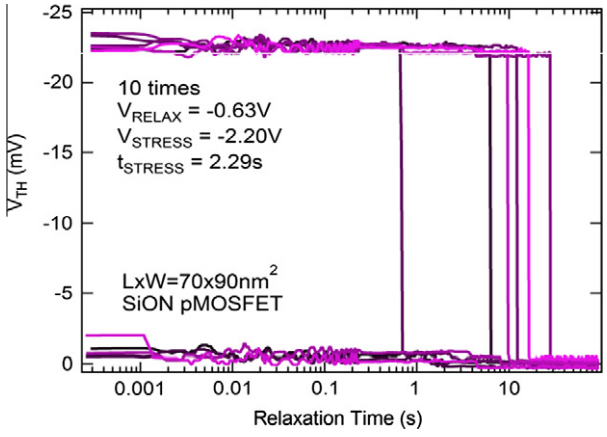

Fig. 4. 10 typical V_{TH} transients registered at −0.63 V after DC stress at −2.2 V for 2.29 s and at 25 °C. 6 out of 10 traces show a giant discrete step of ∼23 mV.

single gate oxide traps in a deeply-scaled FETs during DC and AC bias temperature stress. We demonstrate that the behavior of individual traps as a function of the stress time and duty factor is dictated by their characteristic capture and emission times at high and low voltages at first approximation. A more elaborated model can be found elsewhere [26].

4.1. Emission time and capture times

Fig. 4 shows the relaxation curves at fixed V_{RELAX} after DC stress at V_{STRESS} for a *selected* 90 × 35 nm^2 1.6 nm-*SiON pMOSFET*. A giant V_{TH} shift of ∼23 mV for 6 out of 10 traces at the start of the relaxation period that drops abruptly to 0 mV at t_{RELAX} ∼ 14 s. This step height is significantly larger than the expected threshold voltage shift by the simple charge sheet approximation ($\eta_0 = q/C_{OX}$). As explained previously, this is due to the amplifying effect of the random dopants in the FET channel [11,12,14].

Fig. 5 shows the corresponding TDDS spectra, i.e. two-dimensional histogram of the emission times τ_e and the V_{TH} step heights, obtained from 100 traces of the device under study for two stress times: A homogenous cluster is observed at about 23 mV, indicating the presence of a single active trap. Fig. 6 then shows the histograms of the emission times τ_e (i.e. relaxation time at which the step is detected) obtained at the two stress times. The emission

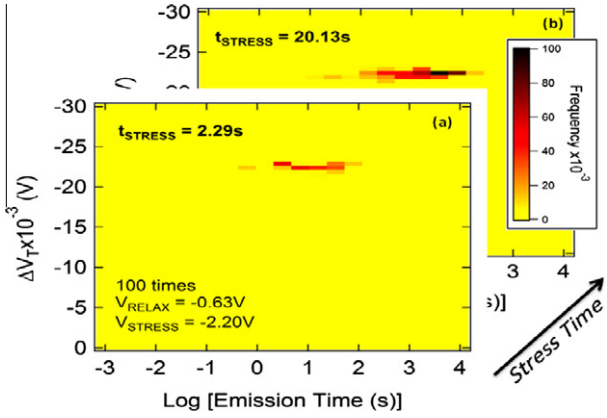

Fig. 5. TDDS spectra for two stress times extracted from 100 recovery traces under the condition of Fig. 4. A homogenous cluster appears at ∼14 s and 23 mV for both spectra. Note that the intensity increases with increasing t_{STRESS} indicating that the trap occupancy after longer stress increases.

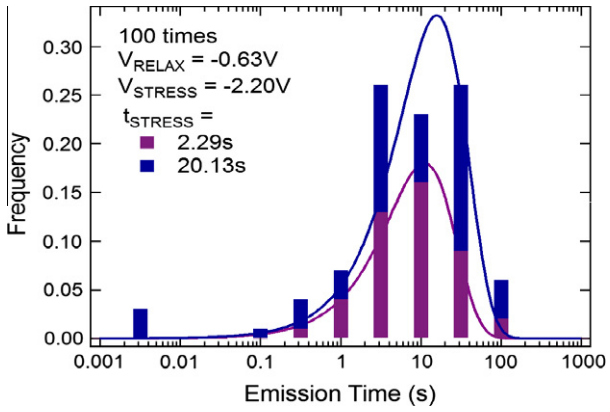

Fig. 6. Histograms f_E of the emission times τ_e extracted from 100 V_{TH} relaxation transients under the condition of Fig. 5. Note that the emission times are binned on the logarithmic scale. The histograms can be fitted with the Eq. (2).

time τ_e follows an exponential distribution as expected from first-order kinetics [18,22]

$$P_E(t_{RELAX}) = \frac{\tau_c}{\tau_c + \tau_e}\left\{1 - \exp\left[-\left(\frac{1}{\tau_e} + \frac{1}{\tau_c}\right)t_{RELAX}\right]\right\}. \quad (1)$$

The emission times can be fitted with the maximum likelihood method and the histogram f_E when binned on logarithmic scale follows

$$f_E(t_{RELAX}) = \frac{t_{RELAX}}{\tau_e}\exp\left[-t_{RELAX}\left(\frac{1}{\tau_e} + \frac{1}{\tau_c}\right)\right], \quad (2)$$

where τ_e and τ_c are the mean emission and capture times at V_{RELAX}, respectively. *Note that the characteristics times have to be always referred to a specific gate voltage*. The fit of the data presented in Fig. 6 provides a characteristic emission time τ_e of about 14 s for both stress times. Therefore, the characteristic emission time is independent of the stress time [6,21]. On the other hand, the characteristic capture time τ_c at V_{RELAX} cannot be fitted accurately since this time is at least one order of magnitude larger than τ_e according to the fit.

The TDDS spectra of Fig. 5 show that the number of traces which present the giant step increases with t_{STRESS}. Fig. 7 shows the intensity of the cluster, i.e. the cumulative probability of charging the trap P_C (=occupancy probability), as a function of t_{STRESS} for two V_{STRESS}. Note that the probability of occupancy saturates at 1 for the highest $|V_{STRESS}|$ and at 0.82 for the lowest $|V_{STRESS}|$. As soon

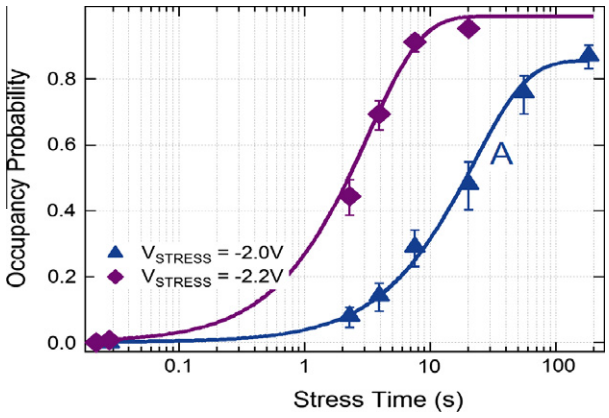

Fig. 7. Trap occupancy probability P_C with $\pm\sigma$ error bars vs. t_{STRESS} for a constant V_{STRESS} of −2.0 and −2.2 V and at 25 °C. For the V_{STRESS} of −2.0 V, the occupancy does not reach 1, indicating that the emission time at V_{STRESS} = −2.0 V is comparable to the capture time.

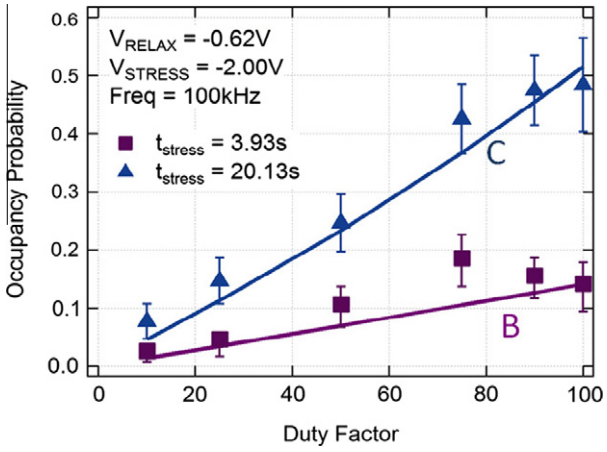

Fig. 8. (Symbols) Trap occupancy probability P_C for V_{STRESS} = −2.00 V and V_{RELAX} = −0.62 V as a function of the AC-stress duty factor DF for two different t_{STRESS} (3.93 and 20.13 s). (Lines) Predicted occupancy according to the proposed model (see Fig. 9 line B for t_{STRESS} = 3.93 s and line C for t_{STRESS} = 20.13 s).

as the emission time enters the same range as the capture time at V_{STRESS}, the probability of intermediate emission during the stress cannot be neglected. All these features can be described by first order kinetics .

$$P_C(t_{STRESS}) = \frac{\tau_e}{\tau_c+\tau_e}\left\{1-\exp\left[-\left(\frac{1}{\tau_e}+\frac{1}{\tau_c}\right)t_{STRESS}\right]\right\}, \quad (3)$$

where τ_e and τ_c are the mean emission and capture times at V_{STRESS}. If τ_c is much shorter than τ_e, the probability of occupancy reaches 1, as shown in Fig. 7 for the largest $|V_{STRESS}|$. In the case of the occupancy saturating at a lower value, the emission events during stress are not negligible and the characteristic emission and capture times can be determined simultaneously from the fit of the data.

In typical BTI experiments, it is observed that the capture time decreases (i.e. the probability of capture increases) and emission time increases with increasing V_{STRESS}. Therefore, at high gate voltages the capture events are dominant, while at low voltages the emission events prevail. For more details about the voltage dependences, consult [6,18,24,27].

4.2. Trap occupancy under AC stress

Fig. 8 shows the occupancy of the trap as a function of the duty factor *DF* of the AC signal for two values of t_{STRESS}. This trend can be explained by considering that the occupancies at both the high/stress (H) and the low/recovery (L) voltages are given by τ_e and τ_c at the corresponding voltage, the times t_H and t_L that each voltage is applied during each period, and the previous state. For the *n-th* period we can thus write

$$P_{CH}(n) = \frac{\tau_{eH}}{\tau_{eH}+\tau_{eH}} + \left\{P_{CL}(n-1) - \frac{\tau_{eH}}{\tau_{eH}+\tau_{eH}}\right\}\exp\left[-\left(\frac{1}{\tau_{eH}}+\frac{1}{\tau_{cH}}\right)t_H\right]. \quad (4)$$

$$P_{CL}(n) = \frac{\tau_{eL}}{\tau_{eL}+\tau_{cL}} + \left\{P_{CH}(n) - \frac{\tau_{eL}}{\tau_{eL}+\tau_{cL}}\right\}\exp\left[-\left(\frac{1}{\tau_{eL}}+\frac{1}{\tau_{cL}}\right)t_L\right]. \quad (5)$$

Expressing the increase in P_{CL} per period, one can obtain the P_C as a function of the number of applied pulses n ($=f\times t_{STRESS}$)

$$P_C(n) = \frac{b}{a}(1-e^{-an}), \quad (6)$$

where a and b are a function of τ_{eH}, τ_{cH}, τ_{eL}, τ_{cL}, *DF*, and f [18,22]. Fig. 9 shows P_C as a function of t_{STRESS} and *DF* calculated using Eq. (6) and using the emission and capture times corresponding to the V_{STRESS} and V_{RELAX} of Figs. 6 and 7. *Note that the lines drawn in Fig. 8 correspond strictly to the model, no fitting is performed.* The proposed model follows correctly the experimental data. This implies that the model can predict correctly the P_C for all t_{STRESS} and *DFs* provided that the emission and capture times are known. Furthermore, *the model can be used to simulate the response of CMOS circuits under AC conditions* as we demonstrate in the last section.

4.3. Temperature activation of the capture and emission times

In order to get insight into the temperature dependence of the emission and capture times, the procedure described in the previous section was applied to the same device at different temperature. The Arrhenius plots, Fig. 10, of the characteristics times provide activation energies from 0.4 eV to 1.0 eV.

Similar thermally activated capture and emission times are also observed in both nFET and pFET with conventional SiO_2 gate oxide [6,20] and high-k dielectrics [8,19]. We therefore conclude for all these cases that *both emission and capture in both electron and hole gate oxide traps are without any doubt thermally activated processes.*

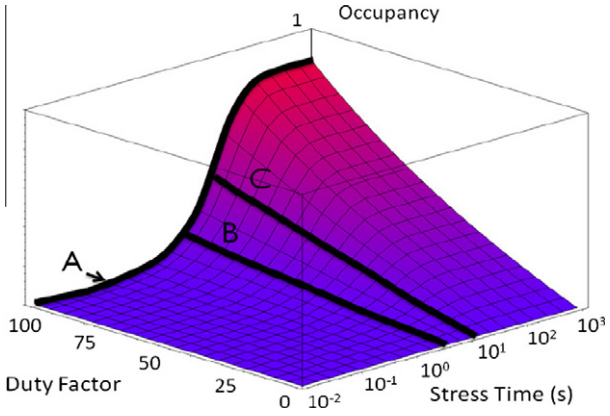

Fig. 9. Occupancy probability as a function of t_{STRESS} and DF considering the conditions of experiment presented in Fig. 6. Line A traces P_C vs. t_{STRESS} at DC stress (see Fig. 7). Lines B and C trace P_C vs. DF (see Fig. 8). The model can predict correctly P_C for all t_{STRESS} and DFs provided that the emission and capture times are known.

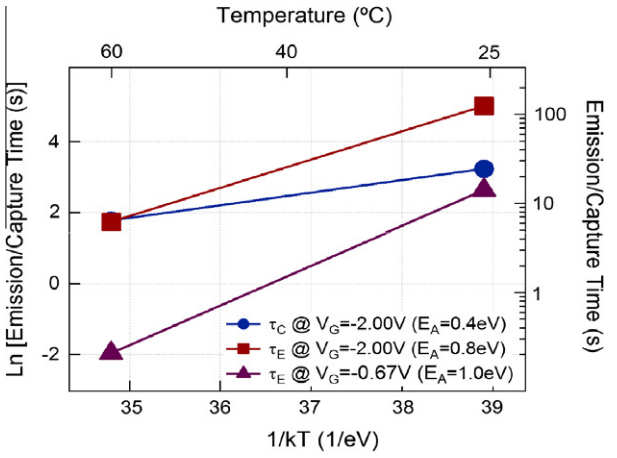

Fig. 10. Arrhenius plot of τ_e and τ_c. Note the large shift of the values for only a 35 °C increase of temperature, indicating that the capture and emission of charge are thermally activated processes [8,11]. The reduction of E_A for τ_e with increasing V_G is in line with the prediction of Ref. [17].

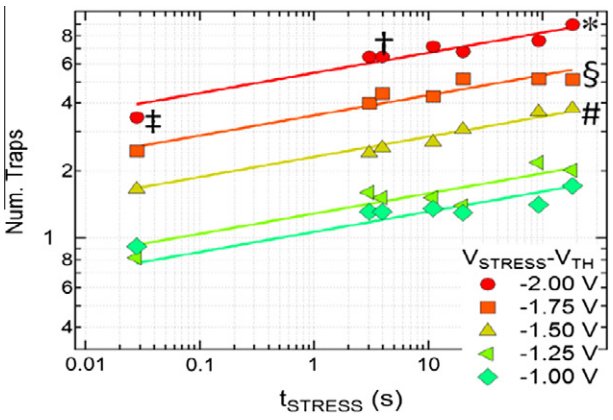

Fig. 12. The number of active traps per device N_T obtained from the fit of the CCDFs shown in Fig. 11 with Eq. (10) (interscept of CCDF with $\Delta V_{TH} = 0$). Note that N_T increases with stress time and voltage. Data can be fitted with both a power-law and logarithmic time dependences [21].

This experimental fact is incompatible with direct elastic tunneling theories widely used in different oxide trap characterization techniques and calculations. Consequently, a new model that takes into account this thermal dependence has to be considered. The most consistent explanation is provided by non-radiative multiphonon (NMP) theory [28] which has recently been applied to BTI data [17].

5. Bias temperature variability

In large devices the random properties of many defects average out resulting in a well-defined lifetime as we showed in Section 2. However, in deeply scaled devices, the stochastic nature of a handful of defects becomes apparent. For this reason, the application of identical workload in such nanoscaled devices results in distributions of the parameter shifts [15,29]. Therefore, the well-defined bias temperature instability (BTI) lifetime of large devices becomes widely distributed [12,22,30]. The atomistic understanding of the properties of individual defects and the demonstrated link between random telegraph noise and BTI presented in the previous section helped us to explain the large BTI variability during relaxation [15,21,31].

In the representative set of quantized NBTI relaxation transients presented in Fig. 2, the *total* ΔV_{TH} (ΔV_{TH} at given t_{RELAX}) strongly varies from device to device. Note that the *total* ΔV_{TH} ranges from a few mV up to 40 mV among devices under identical stress conditions. Fig. 11 shows the complementary cumulative distribution (CCDF = 1 − CDF) of the individual step heights ΔV_{TH} normalized to the number of tested devices. Step heights follow an exponential distribution with an average step height $< \Delta V_{TH} \geq \eta$ equal to 3.4 mV, independent of stress conditions. Note that a single charged defect can cause up to tens of mV of ΔV_{TH}, which is much larger than the value predicted by the simple charge sheet approximation $\eta_0 \sim 1.7$ mV ($\eta/\eta_0 \approx 2.0$ in this case). The number of detected steps increases with stress time and stress voltage (Fig. 11). The number of steps per device is Poisson distributed (figure not shown, Eq. (13)).

The average number of traps per device N_T can be obtained from a maximum likelihood fit of the data with Eq. (10). Fig. 12 shows that N_T follows a power-law voltage dependence and can be fitted with both power law and logarithmic time dependences.

Fig. 13 gives the *total* ΔV_{TH} for pFETs for different (a) t_{STRESS} and (b) V_{STRESS}. The *total* ΔV_{TH} distributions $H_{\eta,NT}$ (ΔV_{TH}) (Eq. (14)) [12,32], a combination of exponential discrete ΔV_{TH} step distributions and the Poisson distributions with average N_T, are traced in Fig. 13 for $\eta = 3.4$ mV and different values of N_T. Note that the lines in Fig. 13 follow the experimental *total* ΔV_{TH} data, and the N_T values given by Eq. (14) excellently match those obtained *independently* in Fig. 11 (see symbols ∗,†,‡,§,#), thus confirming the description derived in Table 1.

A 10 years lifetime CDF prediction of the *total* ΔV_{TH} is obtained by combining Eq. (14) with the N_T dependences on t_{STRESS} and V_{STRESS}. Fig. 14 shows the predicted lifetimes for different conditions. For a fixed failure criteria of $\Delta V_{TH} = 30$, 50, and 100 mV at $t_{STRESS} = 10$ years, Fig. 14a allows one to readily read off the fraction of devices expected to exceed a given failure criterion. As already

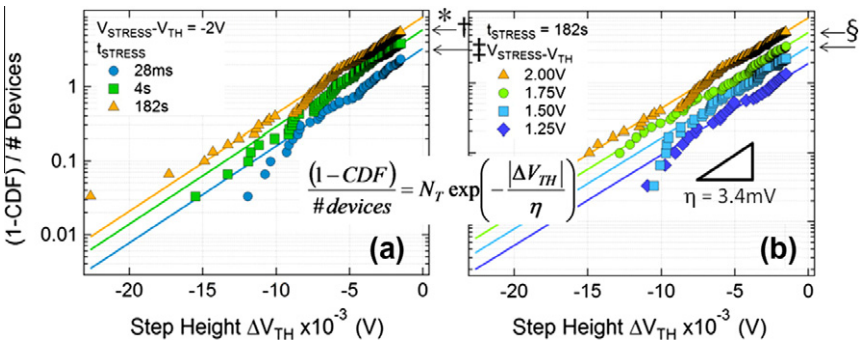

Fig. 11. Complementary cumulative distributions (CCDF = 1 − CDF) of step heights due to single oxide defects normalized to the number of tested pFETs after NBTI follow an exponential distribution (Eq. (10)) with the average step height η. N_T values can be read from the intersection of the fit with the y-axis.

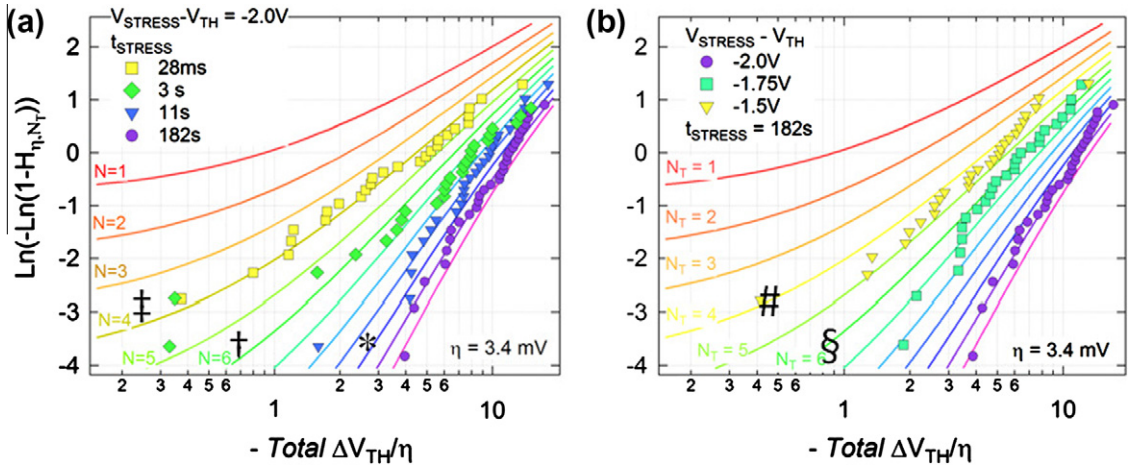

Fig. 13. (Symbols) Cumulative distributions of the total ΔV_{TH} normalized to η = 3.4 mV for 30 pFETs after stress (a) at different voltages and (b) for different times shown in Weibull plots. (Lines) Total ΔV_{TH} CDFs for different N_T values from Eq. (14) match excellently the experimental data.

Table 1
Flow to deduce the total ΔV_{TH} shift distribution [11,21].

Single defect: ΔV_{TH} exponentially distributed with	
$\eta = \langle \Delta V_{TH} \rangle = \frac{\langle x_0 \rangle \sqrt{N_A}}{WL}$	(7)
$f_\eta(\Delta V_{TH}) = \frac{1}{\eta} e^{-\frac{\Delta V_{TH}}{\eta}}$	(8)
$F_\eta(\Delta V_{TH}) = 1 - e^{-\frac{\Delta V_{TH}}{\eta}}$	(9)
$\frac{1-F_\eta(\Delta_{TH})}{\#devices} = N_T e^{-\frac{\Delta V_{TH}}{\eta}}$	(10)
Device: *Total* ΔV_{TH} convolution of n individual exponential distributions = n traps	
$g_{\eta,n}(\Delta V_{TH}) = \frac{\Delta V_{TH}^{(n-1)}}{\eta^n (n-1)!} e^{-\frac{\Delta V_{TH}}{\eta}}$	(11)
$G_{\eta,n}(\Delta V_{TH}) = 1 - \frac{n}{n!} \Gamma(n, \Delta V_{TH}/\eta)$	(12)
Chip: Traps Poisson distributed with $\langle n \rangle = N_T$	
$P_{N_T}(n) = \frac{e^{-N} N_T^n}{n!}$	(13)
Total ΔV_{TH} cumulative distribution in a chip is the sum up of $G_{\eta,n}$ weighted by P_{N_T}	
$H_{\eta,N_T}(\Delta V_{TH}) = \sum_{n=0}^{\infty} \frac{e^{-N} N_T^n}{n!} G_{\eta,n}(\Delta V_{TH})$	(14)

alluded in Section 2, the predicted ΔV_{TH} distribution gets steeper ("tighter") with increasing device area A (Fig. 14b). Since the *average total* ΔV_{TH} is given by $N_T \times \eta$ and $N_T \propto A$, the median *total* $\langle \Delta V_{TH} \rangle$ is independent of A if $\eta \propto 1/A$ [33]. Therefore, Probit $(H_{\eta,N})$ = 0 determines the maximum overdrive for large, i.e., deterministic devices (vertical line in Fig. 14b). In contrast to that a considerable fraction of deeply scaled devices will exceed failure criteria even at low overdrives (see e.g. circles in Fig. 14b). In Fig. 14c and d, the impact of N_T and η on the lifetime is explored. Fig. 14c shows that a reduction of the trap density N_T stretches out the overdrive (horizontal) axis, but the maximum fraction of working devices does not improve significantly at low overdrives. On the other hand, *a reduction of the η value shifts the lifetime prediction vertically*, boosting the number of working devices to high percentages over the whole overdrive range. *The largest gains in reliability can thus be achieved by moving to technologies with reduced dopant concentration N_A in the channel*, see Eq. (7) (Table 1) [3,13,21].

From this study it is evident than a significant fraction of nm-scaled FETs will fail even at low overdrives. This conclusion was anticipated in the introduction and it is obvious considering the link between RTN and BTI, since RTN is a phenomenon that causes giant V_{TH} oscillation at weak inversion, i.e. low overdrives. In future technological nodes, circuit design will become statistical (non deterministic). For this reason, the development of a circuit simulator that accounts for *heteroskedasticity* is compulsory to design reliable circuit with unreliable components.

6. Atomistic approach to BTI variability in circuit simulators

Our "atomistic" simulation framework proposed previously [22,34] is shown in Fig. 15. It allows simulating the impact of *workload*-dependent variability on circuits (i.e., "reliability *distributions* under operating conditions"). The framework accepts the studied circuit in the form of a standard netlist. All or selected FET devices of the input circuit are annotated (i.e., "enhanced") with unique defect properties randomly selected from distributions obtained previously on the simulated technology or from experimental data. These distributions include the (voltage and temperature dependent) capture and emission times [6,30] and the impact of individual defects on the FET properties (e.g., the threshold voltage V_{TH} shift) [12,22]. The latter, as well as the number of defects in each simulated FET device is adjusted to the device gate area. The occupancy of each defect is also determined based on its averaged workload prior to simulated interval in the circuit lifetime [18,22].

Based on this information the control script generates multiple random instances of enhanced circuits and submits them to the HSPICE or SPECTRE solvers. The other crucial component of the framework is the Verilog-A-based BSIM4 FET model augmented to simulate the impact of individual defects on the FET's behavior. It is also capable of following the occupancy of each defect ("defect kinetics") depending on the applied voltages, thus naturally incorporating workload dependence [22,27]. The resulting circuit parameters from all instances are output and subsequently statistically analyzed. The employment of existing industry-standard circuit simulator tools ensures correct combination of the deterministic workload-dependent component with the stochastic modeling aspect while simultaneously incorporating interactions among different devices. The framework has been proven useful for investigation of, e.g., the reliability of SRAMs [35].

7. Conclusions

In this article we have summarized some recent insights into BTI achieved from the comprehensive study of deeply scaled devices. Among the most relevant, it is the close link between RTN and the

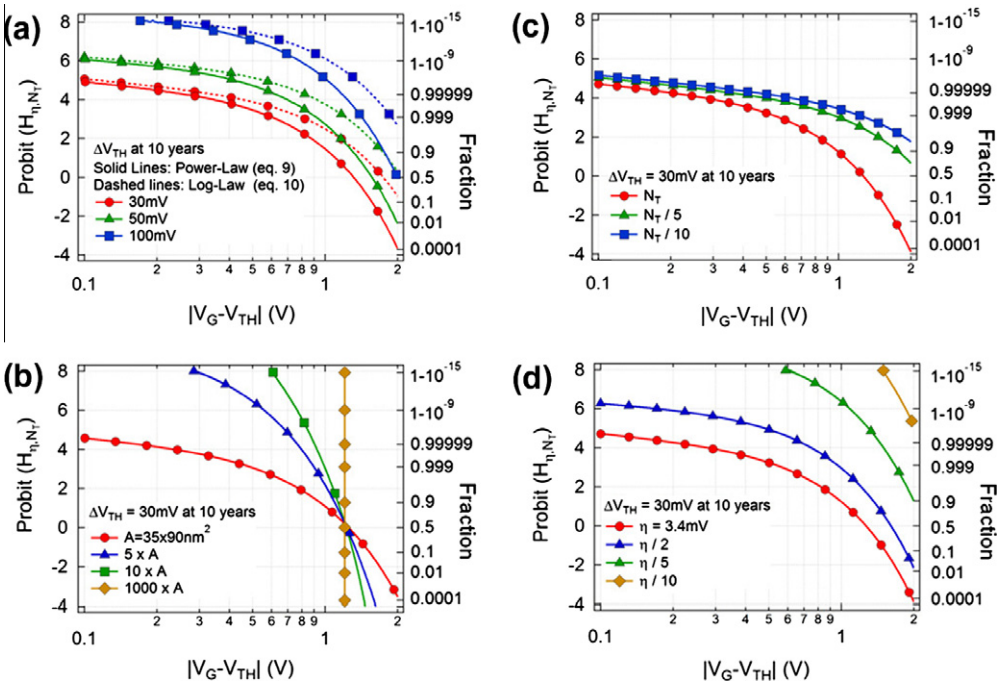

Fig. 14. Predicted 10 years lifetime cumulative distributions of the total pFET ΔV_{TH} at $t_{RELAX} \sim 1$ ms. For different failure criteria (a), a slightly more optimistic prediction is given by a logarithmic time dependent law. For different device areas (b), it is observed that the median total ΔV_{TH} is independent of area. A significant fraction of deeply scaled devices exceeds failure criteria at lower overdrives. For different trap densities (c), CDF stretches out. For different η values (d), a significant boost of the fraction of working devices is obtained.

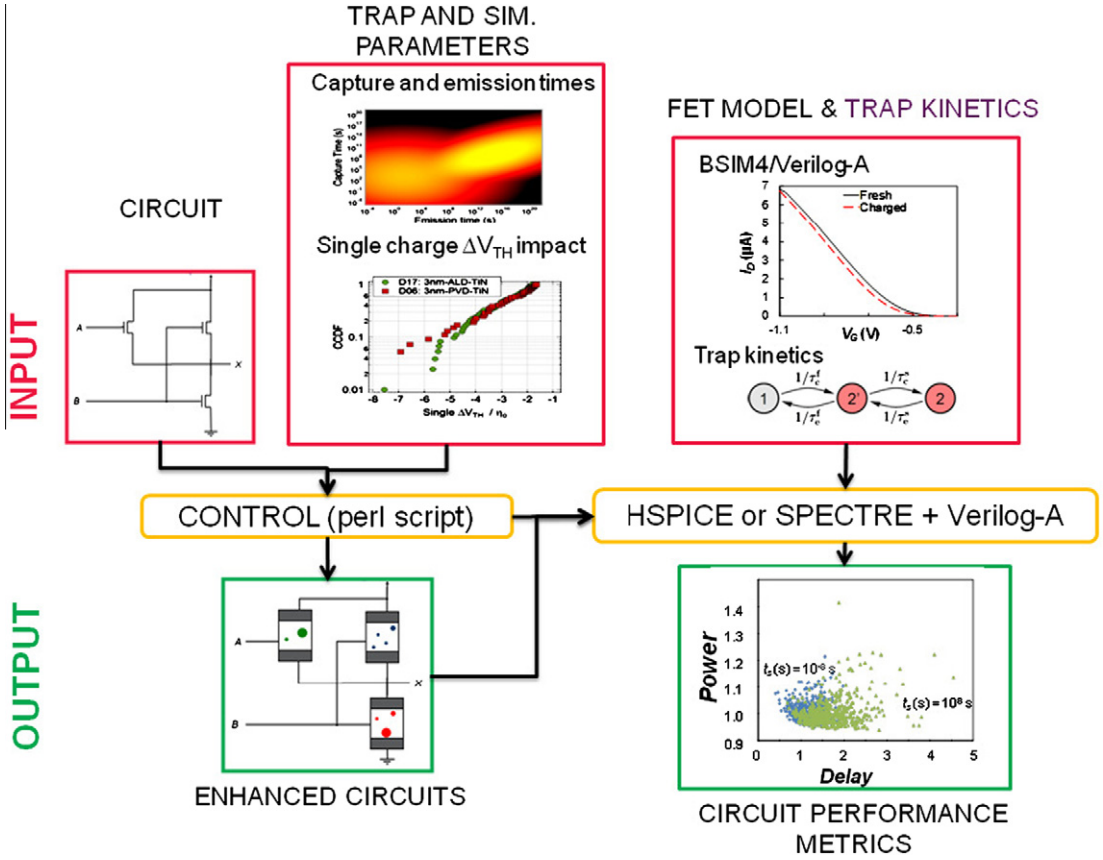

Fig. 15. Simulation setup to study time-dependent variability of circuits based on industry-standard tools.

recoverable component of BTI, indicating that identically behaving traps are responsible of both effects. Useful information about the kinetic properties of individual traps has been straightforwardly extracted from the recently developed technique TDDS. This helped to

understand the charge exchange mechanisms between silicon substrate and gate oxide traps. Based on detailed understanding of the behavior and statistics of individual defects, we have presented a new methodology to predict the BTI lifetime distributions and to develop circuit simulators with deeply scaled FETs.

References

[1] Cartier E, Kerber A, Ando T, Frank MM, Choi K, Krishnan S, et al. Fundamental aspects of HfO2-based high-k metal gate stack reliability and implications on tinv-scaling. In: Proc IEDM; 2011. p. 441–4.

[2] Cho M, Aoulaiche M, Degraeve R, Kaczer B, Franco J, Kauerauf T, et al. Positive and negative bias temperature instability on sub-nanometer EOT high-k MOSFETs. In: Proc IRPS; 2010. p. 1095–8.

[3] Franco J, Kaczer B, Eneman G, Mitard J, Stesmans A, Afanas'ev V, et al. 6Å EOT Si0.45Ge0.55 pMOSFET with optimized reliability (V_{DD} = 1 V): meeting the NBTI lifetime target at ultra-thin EOT. In: Proc IEDM; 2010. p. 70–3.

[4] International Technology Roadmap for Semiconductors. <http://public.itrs.net>.

[5] Kaczer B, Grasser T, Martin-Martinez J, Simoen E, Aoulaiche M, Roussel Ph.J, et al. NBTI from the perspective of defect states with widely distributed time scales. In: Proc IRPS; 2009. p. 55–60.

[6] Grasser T, Reisinger H, Wagner P, Schanovsky F, Goes W, Kaczer B. The time dependent defect spectroscopy (TDDS) for the characterization of the bias temperature instability. In: Proc IRPS; 2010. p. 16–25.

[7] Huard V, Denais M, Parthasarathy C. NBTI degradation: from physical mechanisms to modelling. Microelectron Reliab 2006;46:1–23.

[8] Toledano-Luque M, Kaczer B, Grasser T, Roussel Ph.J, Franco J, Groeseneken G. Toward a streamlined projection of small device BTI lifetime distributions. In: Proc WoDiM; 2012.

[9] Grasser T, Kaczer B, Hehenberger P, Goes W, O'Connor R, Reisinger H, et al. Simultaneous extraction of recoverable and permanent components contributing to bias-temperature instability. In: Proc IEDM; 2007. p. 801–4.

[10] Grasser T, Aichinger Th, Pobegen G, Reisinger H, Wagner P-J, Franco J, et al. The 'permanent' component of NBTI: composition and annealing. In: Proc. IRPS 2011. p. 605–13.

[11] Reisinger H, Grasser T, Gustin W, Schlünder C. The statistical analysis of individual defects constituting NBTI and its implications for modeling DC- and AC-stress. In: Proc IRPS; 2010. p. 7–15.

[12] Kaczer B, Roussel PhJ, Grasser T, Groeseneken G. Statistics of multiple trapped charges in the gate oxide of deeply scaled MOSFET devices—application to NBTI. IEEE Electron Device Lett 2010;31:411–3.

[13] Asenov A, Balasubramaniam R, Brown AR, Davies JH. RTS amplitudes in decananometer mosfets: 3-D simulation study. IEEE Trans Electron Dev 2003;50:839–45.

[14] Bukhori MF, Roy S, Asenov A. Simulation of statistical aspects of charge trapping and related degradation in bulk mosfets in the presence of random discrete dopants. IEEE Trans Electron Dev 2010;57:795–803.

[15] Ghetti A, Compagnoni CM, Spinelli AS, Visconti A. Comprehensive analysis of random telegraph noise instability and its scaling in deca–nanometer flash memories. IEEE Trans Electron Dev 2009;56:1746–52.

[16] Kaczer B, Grasser1 T, Roussel Ph.J, Franco J, Degraeve R, Ragnarsson L-A, et al. in Proc IRPS; 2010. p. 26–32.

[17] Kaczer B, Grasser T, Roussel PJ, Martin-Martinez J, O'Connor R, O'Sullivan BJ, et al. Ubiquitous relaxation in BTI stressing—new evaluation and insights. In: Proc.: IRPS 2008. p. 20–7.

[18] Grasser T, Reisinger H, Wagner P-J, Kaczer B. Time-dependent defect spectroscopy for characterization of border traps in metal–oxide–semiconductor transistors. Phys Rev B 2010;82:245318.

[19] Toledano-Luque M, Kaczer B, Roussel Ph.J, Grasser T, Wirth GI, Franco J, et al. Response of a single trap to AC negative bias temperature stress. In: Proc IRPS; 2011.

[20] Toledano-Luque M, Kaczer B, Simoen E, Roussel PhJ, Veloso A, Grasser T, et al. Temperature and voltage dependences of the capture and emission times of individual traps in high-k dielectrics. Microelectron Eng 2011;88:1243–6.

[21] Toledano-Luque M, Kaczer B, Roussel Ph, Cho MJ, Grasser T, Groeseneken G. Temperature dependence of the emission and capture times of sion individual traps after positive bias temperature stress. J Vac Sci Technol B 2011;29. 01AA04.

[22] Toledano-Luque M, Kaczer B, Franco J, Roussel Ph.J, Grasser T, Hoffmann TY, et al. From mean values to distributions of BTI lifetime of deeply scaled FETs through atomistic understanding of the degradation. In: Proc VLSI; 2011.

[23] Kaczer B, Mahato S, Valduga de Almeida Camargo V, Toledano Luque M, Roussel Ph.J, Grasser T, et al. Atomistic approach to variability of bias-temperature instability in circuit simulations. In: Proc IRPS; 2011. p. 915–9.

[24] Chia-Yu C, Qiushi R, Hyun-Jin C, Kerber A, Yang L, Ming-Ren L, et al. Correlation of Id- and Ig-random telegraph noise to positive bias temperature instability in scaled high-κ/metal gate n-type MOSFETs. In: Proc IRPS; 2011. p. 190–5.

[25] Toledano-Luque M, Kaczer B, Simoen E, Degraeve R, Franco J, Roussel Ph.J, et al. Correlation of single trapping and detrapping effects in drain and gate currents of nanoscaled nFETs and pFETs. In: Proc IRPS; 2012. p. XT.5.1–6.

[26] Grasser T, Kaczer B, Reisinger H, Wagner P-J, Toledano-Luque M. On the frequency dependence of the bias temperature instability. In: Proc IRPS; 2012. p. XT.8.1–7.

[27] Grasser T. Stochastic charge trapping in oxides: from random telegraph noise to bias temperature instabilities. Microelectron Reliab 2011;52:39–70.

[28] Uren M, Kirton M, Collins S. Anomalous telegraph noise in small-area silicon metal–oxide–semiconductor field-effect-transistors. Phys Rev B 1988;37:8346–50.

[29] Huard V, Parthasarathy C, Guerin C, Valentin T, Pion E, Mammasse M, et al. NBTI degradation: from transistor to SRAM arrays. In: Proc IRPS; 2008. p. 289–300.

[30] Grasser T, Wagner P-J, Reisinger H, Aichinger Th, Pobegen G, Nelhiebel M, et al. Analytic modeling of the bias temperature instability using capture/emission time maps. In: Proc IEDM; 2011. p. 6618–21.

[31] Grasser T, Kaczer B, Goes W, Reisinger H, Aichinger Th, Hehenberger Ph, et al. Recent advances in understanding the bias temperature instability. In: Proc IEDM; 2010. p. 82–5.

[32] Takeuchi K, Nagumo T, Yokogawa S, Imai K, Hayashi Y. Single-charge-based modeling of transistor characteristics fluctuations based on statistical measurement of RTN amplitude. In: Proc VLSI; 2009. p. 54–5.

[33] Franco J, Kaczer B, Toledano-Luque M, Roussel PhJ, Mitard J, Ragnarsson L-Å, et al. Impact of single charged gate oxide defects on the performance and scaling of nanoscaled fets. IEEE Electron Dev Lett 2012;33:779–81.

[34] Kaczer B, Franco J, Toledano-Luque M, Roussel Ph.J, Bukhori MF, Asenov A, et al. The relevance of deeply-scaled FET threshold voltage shifts for operation lifetimes. In: Proc IRPS; 2012. p. 5A.2.1–6.

[35] Rodopoulos D, Mahato SB, de Almeida Camargo VV, Kaczer B, Catthoor F, Cosemans S, et al. Time and workload dependent device variability in circuit simulations. In: Proc ICICDT; 2011.