Microelectronics Reliability 81 (2018) 186–194

Contents lists available at ScienceDirect

Microelectronics Reliability

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

Invited paper

A brief overview of gate oxide defect properties and their relation to MOSFET instabilities and device and circuit time-dependent variability

B. Kaczer[a,*], J. Franco[a], P. Weckx[a], Ph.J. Roussel[a], V. Putcha[a,b], E. Bury[a], M. Simicic[a,b], A. Chasin[a], D. Linten[a], B. Parvais[a], F. Catthoor[a,b], G. Rzepa[c], M. Waltl[c], T. Grasser[c]

[a] imec, Kapeldreef 75, B-3001 Leuven, Belgium
[b] KU Leuven, ESAT Department, B-3001 Leuven, Belgium
[c] TU Wien, Gusshausstrasse 27-29, A-1040 Vienna, Austria

ARTICLE INFO

Keywords:
Bias Temperature Instability (BTI)
Random Telegraph Noise (RTN)
Gate oxide defects
Variability
Circuit simulations

ABSTRACT

A paradigm for MOSFET instabilities is outlined based on gate oxide traps and the detailed understanding of their properties. A model with trap energy levels in the gate dielectric and their misalignment with the channel Fermi level is described, offering the most successful strategy to reduce both Positive and Negative Bias Temperature Instability (PBTI and NBTI) in a range of gate stacks. Trap temporal properties are determined by tunneling between the carrier reservoir and the trap itself, as well as thermal barriers related to atomic reconfiguration. Trap electrostatic impact depends on the gate voltage and its spatial position, randomized by variations in the channel potential. All internal properties of traps are distributed, resulting in distributions of the externally observable trap parameters, and in turn in time-dependent variability in devices and circuits.

1. Introduction

Electrically active gate oxide defects are an inevitable reality of MOS-based devices. Indeed, their control and reduction in silicon dioxide was one of the main enablers of silicon-based microelectronics [1,2]. The excess of gate oxide defects is also the main roadblock for post-Si MOS technologies, such as those based on Ge and IIIV channels [3]. Gate oxide defects are responsible for sub-optimal FET parameters of as-fabricated devices, such as mobility, sub-threshold slope, and threshold voltage, as well as a further degradation of these parameters during device operation. Depending on the type of measurement and the measurement conditions, these time-dependent instabilities are commonly referred to as hysteresis, Bias Temperature Instability (BTI) and Random Telegraph Noise (RTN), the latter observable in smaller-area devices.

Negative BTI (NBTI), observed since the early period of MOSFET technology, remains a critical issue in present-day Si-based p-channel MOSFETs [4,5]. With the introduction of high-k gate dielectrics, Positive BTI (PBTI) emerged as a concern for n-channel MOSFETs, later on minimized by high-k material quality improvement; nevertheless, PBTI still represents a concern for beyond-Si technologies [3,6].

In general, the instabilities are caused by charging and discharging of *as-fabricated* and *generated* bulk and interface states [7–9]. Over the last decade we have concluded that a large body of observations in a range of Si and post-Si gate stacks can be explained if the detailed properties of the gate oxide defects are understood and invoked. Primarily, the trap *energy levels* in the gate dielectric and their *misalignment* with the channel Fermi level offer a picture that can explain many features of both PBTI and NBTI in a range of stacks. The channel/trap energy level misalignment is also proposed as the most efficient method to reduce BTI, particularly in post-Si gate stacks, and is discussed in the next section, i.e., Section 2.

The rapid downscaling of FET devices brought about the possibility to observe *individual* trapping and detrapping events and with it, direct experimental access to additional trap properties. These include the *internal barriers* associated with the trap atomic reconfiguration (aka *structural relaxation*) during carrier capture and emission, which are e.g. responsible for the strong temperature dependence of the capture and emission times. A closer investigation then revealed even a more intricate, multi-state structure of some traps, allowing to explain complex RTN behavior, such as anomalous RTN and capture time frequency dependence [10,11].

Apart from their temporal properties, the electrostatics of single traps could be investigated in deeply scaled devices. The effect of single traps on the FET characteristics depends on the trap depth in the oxide, lateral position above the channel, as well as random variations in the FET channel potential [12]. The main properties of individual defects are reviewed in Section 3.

* Corresponding author.
E-mail address: kaczer@imec.be (B. Kaczer).

https://doi.org/10.1016/j.microrel.2017.11.022
Received 17 November 2017; Accepted 25 November 2017
Available online 04 January 2018
0026-2714/

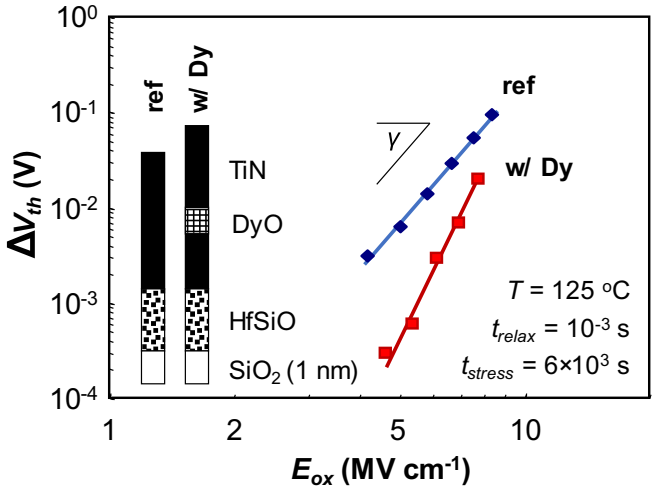

Fig. 1. A significant reduction of PBTI ΔV_{th} observed in nFinFETs with Dy (stack "w/ Dy", see inset) over the reference stack ("ref") is ascribed to Dy diffusion into the HfSiO dielectric layer, resulting in trap "passivation" by energy level shifting. Note the accompanying "signature" increase in the ΔV_{th} oxide electric field (~gate voltage) dependence E_{ox}^{γ}. A ~9% increase in equivalent oxide thickness of the "Dy" sample alone cannot explain the improvement.

Decomposing transistor gate oxide reliability down into individual traps and their well understood properties allows us to reassess and reassemble the time-dependent behavior of deeply scaled devices from "bottom up" [13]. The collective action of several traps in each device results in within-device and device-to-device time-dependent variability [9]. The methods of incorporating this variability in reliability-aware circuit simulations, the practical application of our understanding, are briefly mentioned in Section 4.

2. Trap level perspective of BTI

Positive BTI (PBTI) came into prominence with the advent of high-k gate oxides [14]. We and others have noticed that the inclusion of rare earth elements in the high-k layer results in the reduction of the threshold voltage shift ΔV_{th} after a PBTI stress, Fig. 1 [15,16]. Remarkably, the instability *reduction* is accompanied by a *larger* oxide electric field (or gate voltage) dependence. A larger field (or voltage) power-law exponent γ (i.e., the slope in a log-log plot) for the stack with dysprosium implies a larger reduction of instability (i.e., higher reliability) at lower, operating fields or voltages.

We have observed similar trends in p-channel FETs with Si-capped SiGe and Ge channels, considered for beyond 10 nm VLSI nodes [17,18]. One can again see a major reduction in *Negative* BTI (NBTI) instability in Si-capped Ge-based devices, Fig. 2a, accompanied by an increase in voltage acceleration at lower biases. NBTI reduction is also observed when the SiGe quantum well thickness is increased, Fig. 2b. Most intriguingly, NBTI is reduced when the thickness of the Si capping layer decreases, Fig. 2c. This last observation is the most counter-intuitive, as reducing the Si cap brings the Si/SiO_2 interface and the gate oxide traps spatially closer to the inversion layer holes, so a higher degradation would be naively expected.

Based on these observations for both PBTI and NBTI, Figs. 1 and 2, respectively, we have proposed a *unified model* for BTI trapping, based on *channel carrier/trap energy level misalignment*, see Fig. 3 [18]. In the case of PBTI, when the energy level of electron traps is shifted upward, toward the gate oxide conduction band, Fig. 3a, the traps become less accessible for n-channel electrons. Analogously for NBTI in the p-channel stacks, the introduction of i) lower-band gap (Si)Ge, with different valence band alignment to SiO_2, ii) a wider (Si)Ge quantum well with narrower sub-bands, and iii) a thinner Si cap, Fig. 2a, b, and c, respectively, results in misaligning the channel holes with gate oxide

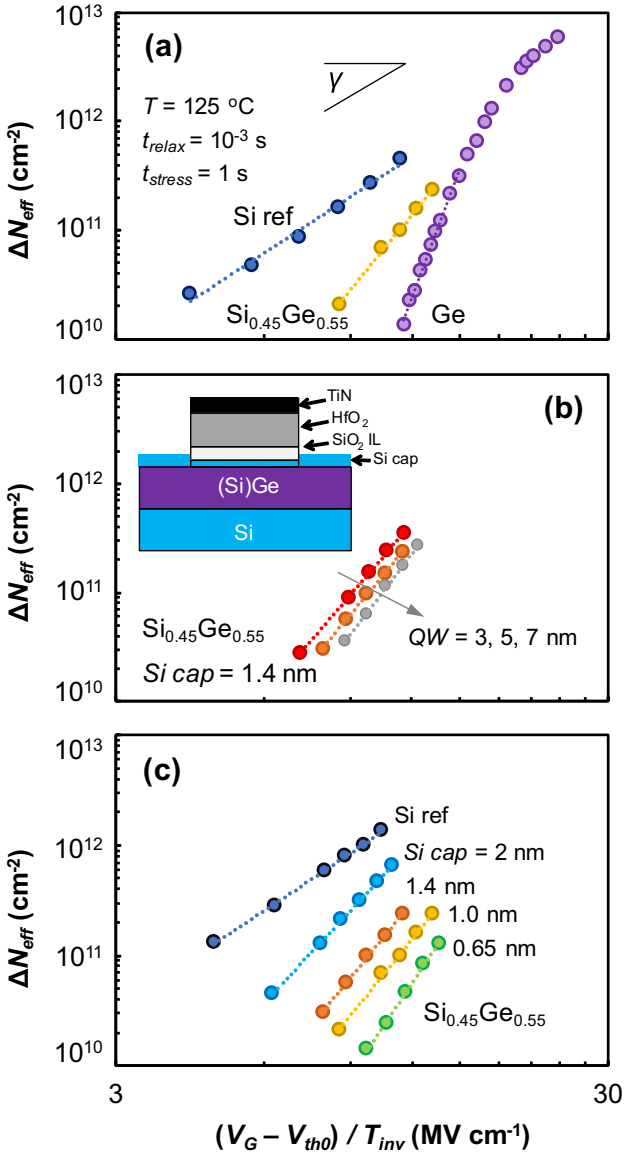

Fig. 2. Plot of $\Delta N_{eff} = C_{ox}\,\Delta V_{th} / q$ vs. the oxide electric field *estimate* $(V_G - V_{th0}) / T_{inv}$ allows comparing BTI in disparate gate stacks. (a) A higher Ge fraction in the channel of a pFET, see the inset of (b), lowers ΔN_{eff}, i.e., boosts NBTI robustness. (b) The same effect is observed as (b) SiGe quantum well thickness (QW) increases and (c) Si cap thickness decreases. In all three panels, note the voltage power-law exponent γ (slope in a log-log plot) increases in stacks with lower overall ΔN_{eff} at operating voltages.

trap levels and decreased NBTI, Fig. 3b.

We note that our model for BTI trapping is in fact merely repurposing established notions about charge trapping in MOS layers [11,19–21]. In our opinion, the trap-level picture of BTI should prove even more useful going forward, as novel substrates and SiO_2-free oxide combinations are considered and studied [3]. For example, the trap-level model excellently explains multiple properties of I-V hysteresis in $InGaAs/Al_2O_3$ and $InGaAs/HfO_2$ gate stacks, Fig. 4 [3].

Understanding BTI in terms of oxide trap energy levels is intuitive and inherently enables multiple options to reduce BTI by gate stack design and processing. Trap level shifting can be achieved for example by rare-earth or even nitrogen "doping" of HfO_2 (for electron traps) [22–24], dipole engineering [25,26], and dedicated anneals [27,28].

In the framework of the proposed picture we can argue the following. Reducing the number of gate oxide defects will result in a proportional BTI reduction *uniformly at all gate voltages*. In contrast, misaligning the trap levels, without reducing the trap density, will

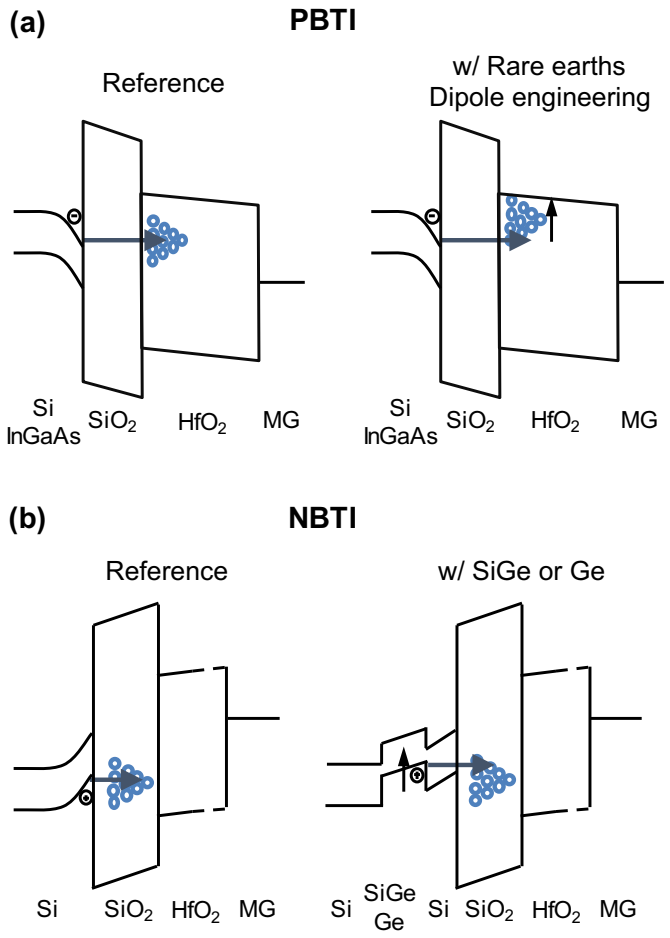

Fig. 3. A schematic illustration of the reduction of charge trapping by decoupling defect and channel energy levels in (a) nFETs (PBTI), by introducing "doping" and dipole-forming elements into the high-k dielectric layer, and in (b) pFETs (NBTI), by introducing a low-bandgap Ge channel with valence band offset with respect to Si.

result in *drastic improvements at low-bias operating conditions*, Fig. 5 [29]. This is because in the latter case, channel carriers can access only the tail of the shifted trap level distribution, Fig. 3. Note also that, contrary to the simple trap density reduction, trap level misalignment naturally explains the increased BTI voltage acceleration, i.e., the increased power-law exponent γ, cf. Fig. 5 and Figs. 1 and 2.

3. Oxide trap properties

When the stress bias is removed from the FET gate, BTI degradation starts to recover immediately. Especially ΔV_{th} recovery (aka *relaxation*) is archetypal for the BTI measurement. It has been the source of most of the controversy surrounding the BTI phenomenon, also because it complicates comparison of results if not duly measured and specified.

The ΔV_{th} relaxation in large devices is log-like and typically featureless, Fig. 6a [30]. It is observed to already be in progress at 1 μs or less after stress removal and continues for up to many days [31,32]. We have long argued that the BTI recovery contains key information about the underlying physical mechanisms and should therefore be correctly included in a valid BTI model [33]. Namely, the log-like behavior spanning many decades indicates the *relaxation process is composed of a wide range of characteristic time scales*.

When the same relaxation measurement is performed on a small device (gate area $< 10^4$ nm^2), individual discrete steps appear in the relaxation trace, Fig. 6a. As in a static RTN measurement, *each of these steps corresponds to an individual detrapping event in a single gate oxide trap* [13,34,35].

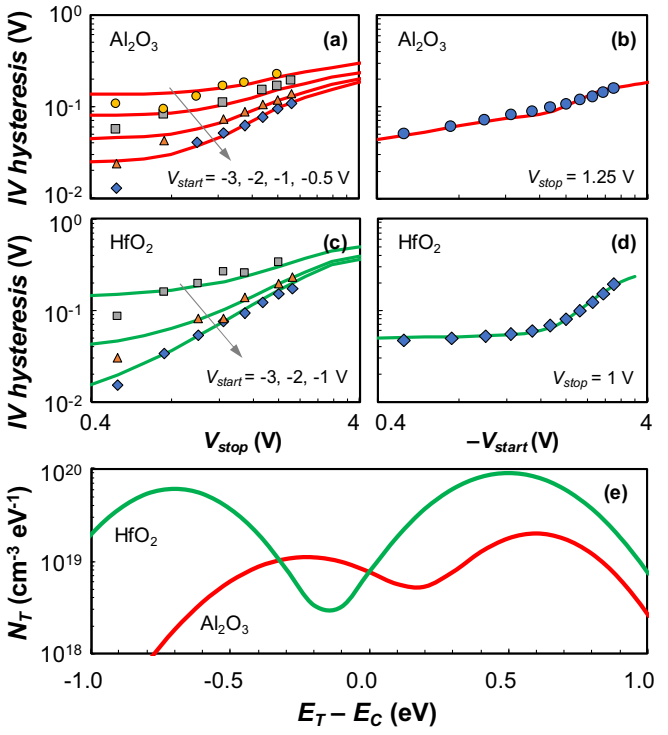

Fig. 4. Measured I-V hysteresis for varying V_{stop} and V_{start}, in (a–b) InGaAs/5 nm Al_2O_3, and (c–d) InGaAs/5 nm HfO_2. (e) A defect band model, including two normal distributions of defect levels above and below the InGaAs E_C, calibrated to match the experimental data, cf. lines in (a–d). Al_2O_3 shows lower peak defect densities, but defect bands overlap, inducing an approx. uniform distribution around InGaAs E_C. In contrast, HfO_2 shows a minimum defect density ~0.2 eV below the channel E_C.

Furthermore, upon repeated stress and remeasurement of the relaxation trace, the same steps reappear *stochastically* with a probability distribution around their characteristic relaxation, i.e., mean emission time $\langle\tau_e\rangle$, inset of Fig. 6a. This probability distribution is observed to be exponential in time and *independent of the stress time*. This indicates a reaction-controlled trapping that can be described by a Markov process. This is in contrast to predictions made by the reaction-diffusion model, in which the $\langle\tau_e\rangle$ would increase with increasing stress time and the τ_e distribution would be log-normal [36].

The repeated stress-and-relaxation measurement on a small device is the basis of the so-called Time-Dependent Defect Spectroscopy Technique (TDDS) [35]. The step heights and the distributed emission times from multiple traps then appear as clusters in the TDDS spectrum, Fig. 6b. The spectrum varies from device to device and constitutes a unique fingerprint of defects *active* in the particular device [37].

Because of its active perturbation, i.e., the "stress" phase, the TDDS measurement is akin to dynamic or cyclo-static RTN measurements [38]. TDDS therefore typically reveals a larger number of traps in a device than would be visible in a static RTN measurement. Through changing the perturbation duration, amplitude or shape, other trap parameters, such its occupancy and mean capture time $\langle\tau_c\rangle$ can be measured [39–41]. As in the case of emission, individual *capture* events are exponentially distributed in time around $\langle\tau_c\rangle$ [40].

Static RTN and TDDS measurements reveal that both capture and emission times of individual traps are typically strongly temperature activated, Fig. 7 [40,42]. The activation energies, moreover, vary substantially from trap to trap [43]. Trap capture $\langle\tau_c\rangle$ and emission $\langle\tau_e\rangle$ times are therefore determined by i) the tunneling distance of carriers into/out of the traps and ii) a thermally activated and bias-independent component, Fig. 8. The presence of the latter, widely-distributed process can explain the very long τ_c and τ_e times observed in some traps even in very thin gate oxides [36,40]. On the other hand, traps with

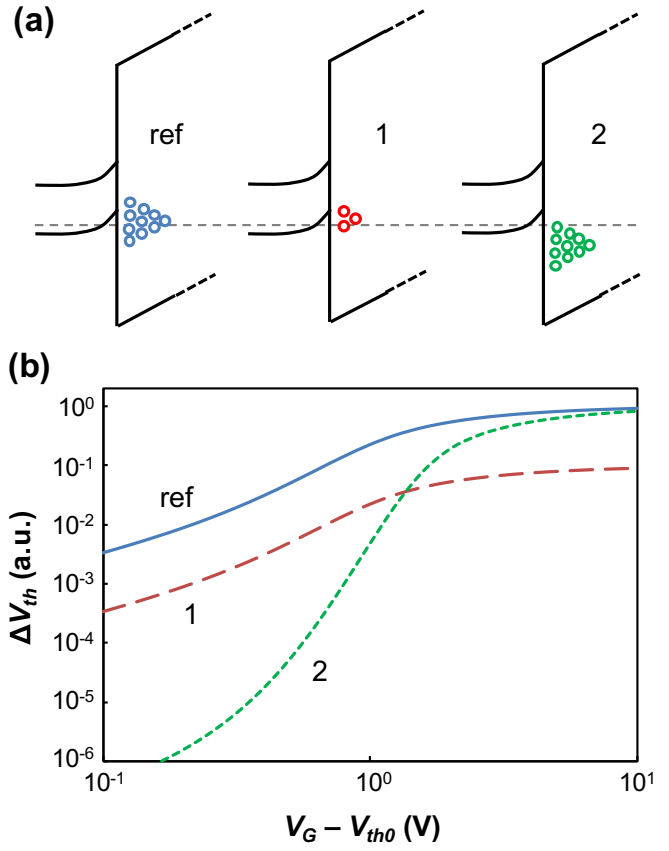

Fig. 5. (a) With respect to the reference case (ref), charge trapping is suppressed by i) reducing the dielectric defect density (Scenario 1), or ii) by carrier/defect energy decoupling (Scenario 2). (b) Calculated ΔV_{th} assuming a 10× defect density reduction (Scenario 1), or the reference defect density of states but with the mean shifted by 0.5 eV (Scenario 2). The latter case yields an exponential reduction of charge-trapping at operating V_G. Note also that Scenario 2 reproduces the observation of increased power-law exponent γ in Figs. 1 and 2.

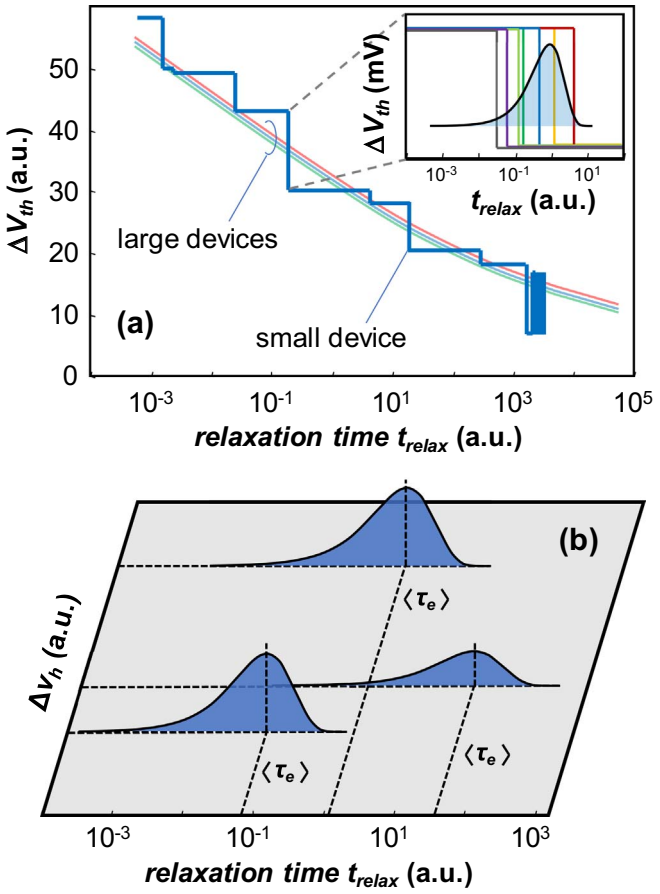

Fig. 6. (a) ΔV_{th} relaxation in *large* devices is generally featureless, log-like and distributed over a wide range of time scales. All large devices behave nearly identically. In contrast, multiple *discrete* detrapping events are visible in *small* device relaxation traces. Inset: upon repeated stress-and-relaxation measurement of a single small device, the same Δv_{th} step corresponding to the same detrapping event is observed with exponentially distributed instances of emission times τ_e. (b) Emission times τ_e, together with the corresponding Δv_{th} values, extracted from the repeated relaxation traces can be plotted in parallel for multiple traps into a 2D histogram, forming a TDDS spectrum of the device. Distribution magnitudes reflect respective trap occupancies.

very short τ_c and τ_e times ($< 10^{-3}$ s) will be typically interpreted in standard CV and Charge Pumping measurements as interface traps [44].

The thermally activated process is identified as a reconfiguration, or a *structural relaxation* of the atoms surrounding the trapping site during (dis-) charging. The total energy of the system during the (dis-) charging process is conserved through capture or emission of multiple phonons [11,45–49]. The energy barriers $\varepsilon_{12}(V_G)$ and $\varepsilon_{21}(V_G)$, associated respectively with capture and emission thermal activation are tied to the Huang-Rhys factor S, Fig. 8, and parabolic curvature ratio R (not shown) [11]. To account for the observed variations in the thermal activation of $\langle\tau_c\rangle$ and $\langle\tau_e\rangle$, factor S, Fig. 8, and optionally, R, should be assumed to be distributed [27].

Capture and emission times $\langle\tau_c\rangle$ and $\langle\tau_e\rangle$ of each trap are therefore determined by its i) energy level E_T (cf. Section 2), i) trap depth z_T, and, for a 2-state system, iii) parameters S and R. Note that τ_c and τ_e depend on the gate voltage V_G through its leveraging action on ΔE_{12}, which in turn affects barriers $\varepsilon_{12}(V_G)$ and $\varepsilon_{21}(V_G)$, Fig. 8. The local availability of carriers in the reservoir should be also considered [50].

Finally, from the *distributions* of the above three factors, summarized in Fig. 8, the *distributions* of $\langle\tau_c\rangle$ and $\langle\tau_e\rangle$ and their V_G and T dependences can be calculated, Fig. 9 [51]. As will be discussed below, the $\langle\tau_c\rangle(V_G, T)$ and $\langle\tau_e\rangle(V_G, T)$ distributions will prove useful for trap-enhanced circuit simulations.

A closer examination of possible trap configurations reveals the internal structure of some traps to be more complex than the basic 2 states (i.e., charged and discharged). The additional meta-stable states, Fig. 8, are then argued to be responsible for second-order effects, such as switching oxide traps, anomalous RTN, multistate frequency dependence, and bias independent time constants, as well as gate leakage [10,11,52,53].

Let us now turn our attention from the trap temporal properties to their impact on the FET electrostatics. In deeply scaled devices with lateral dimensions comparable to the Debye length, the potential distribution in the channel cannot be considered smooth but modulated by individual charged dopants and other sources of variability. Consequently, the source-to-drain conduction proceeds in a non-uniform, percolative manner, Fig. 10a [12,54].

A single charged trap in the gate oxide will locally affect the channel potential through a Coulombic interaction. The impact of *single* charged traps on the channel current, represented by the threshold voltage shift Δv_{th},[1] varies vastly, both in magnitude and in V_G dependence, Fig. 10b. To a first-order approximation, we have found it to depend on the lateral location of the trap with respect to the current percolation paths—traps closer to the paths, and specifically, percolation constriction spots, will have a significant impact. Conversely, charging of traps acting over a region with little channel current will be barely noticed at

[1] v_{th} is used to represent the effect of an *individual* trap, while V_{th} represents the *total* FET threshold voltage shift due to *multiple* traps.

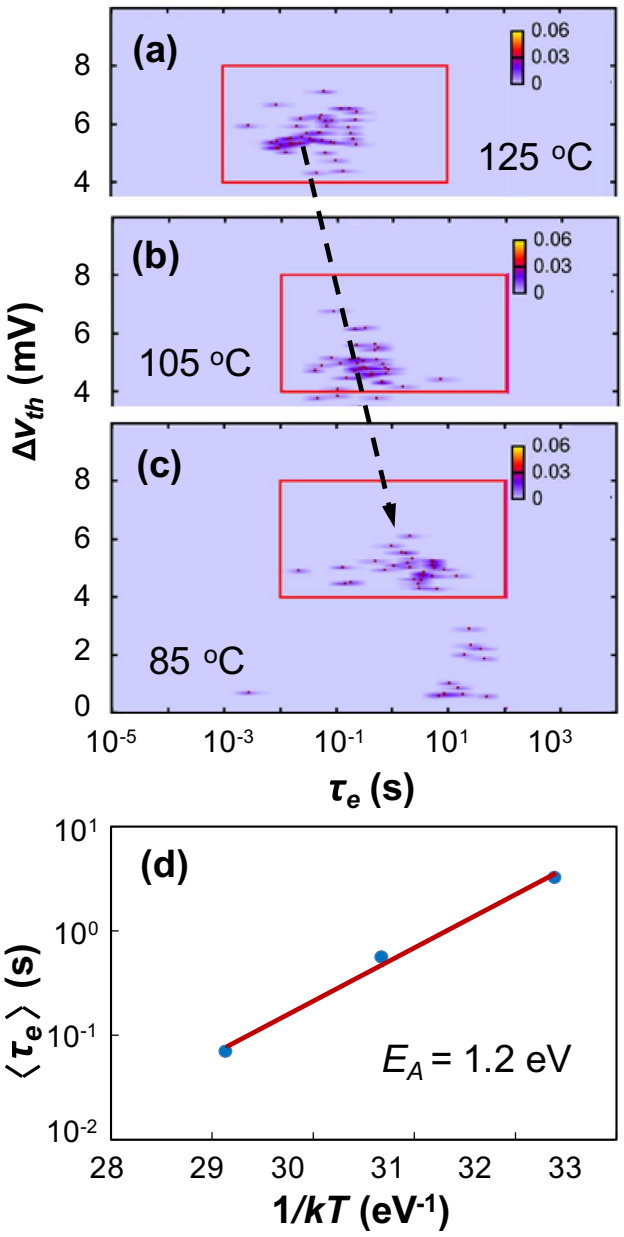

Fig. 7. (a–c) An example of TDDS spectra with one prominent electron trap emission recorded at 3 temperatures on a commercial 28 nm high-k technology. The trap emission time τ_e is observed to be strongly thermally activated, E_A = 1.2 eV in this particular case (d).

the drain terminal [12,50,55].

The distribution of individual trap impacts Δv_{th} can now be understood. Firstly, traps near the junctions will have a smaller impact on the current and hence Δv_{th}, Fig. 11a. As discussed in the previous paragraph, random channel variations will further randomize this impact. We have found that this stochastic component appears to be log-*normally* distributed, Fig. 11b, with the shape parameter σ_{perc} increasing approximately with the square root of the channel doping [56,57]. The resulting distribution of Δv_{th} is therefore a *convolution* of the trap mean and the percolative effects, Fig. 11c. For intermediate values of σ_{perc}, corresponding to typical channel doping levels, the single-trap Δv_{th} distribution is approximately exponential, in line with experimental observations [58]. The expectation value of the exponential distribution is $\eta \equiv \langle \Delta v_{th} \rangle$, i.e., the mean impact of a single charged trap on the transistor's threshold voltage.

Simulations show that the mean trap impact η is approximately

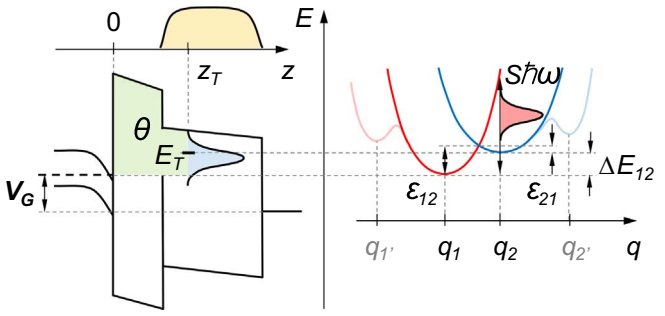

Fig. 8. In general, gate oxide traps are distributed in i) depth z_T ("yellow" distribution), ii) energy E_T ("blue" distribution), and iii) internal configuration, represented here for a 2-state case by the Huang-Rhys factor S ("red" distribution) in the configuration coordinate q. The capture (emission) time of a trap $\Delta E_{12}(V_G)$ above the channel Fermi level is therefore determined by i) the tunneling barrier θ (green) between the carrier reservoir (nFET inversion layer here) and the trap and ii) the barriers $\varepsilon_{12}(V_G)$ and $\varepsilon_{21}(V_G)$ between the unoccupied (1) and occupied (2) states. Due to ii) the transition is typically strongly temperature dependent. Possible meta-stable states (1′) and (2′) are responsible for a more complex trap behavior. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

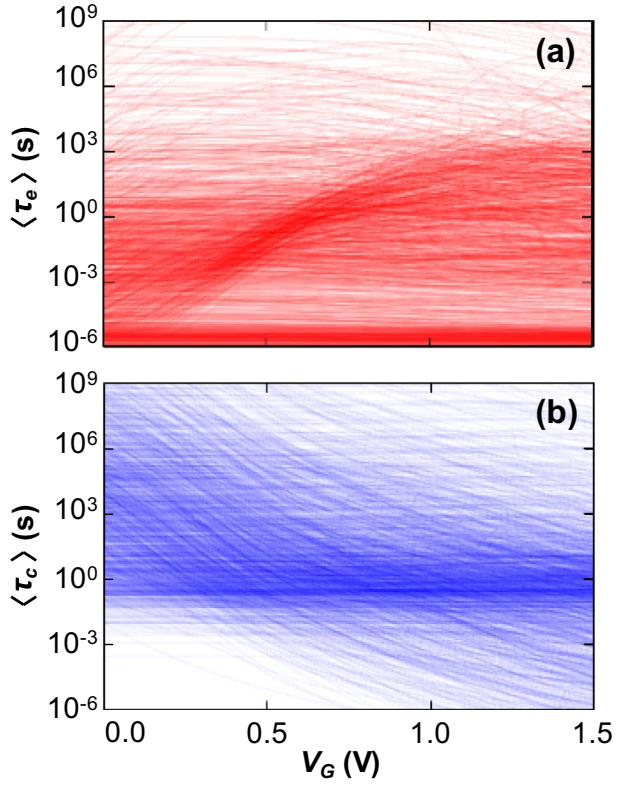

Fig. 9. *Distributions* of (a) capture time $\langle \tau_c \rangle(V_G)$ and (b) emission time $\langle \tau_e \rangle(V_G)$ gate-voltage dependences of a population of defects in a pFET gate oxide at a given temperature, calculated so that the combined effect of traps fits large-area pFET eMSM measurements.

proportional to σ_{perc}, i.e. $\eta \propto \sigma_{perc}$ increases roughly as a square root of channel doping [57]. However, in the absence of dopant-induced variation of the channel potential, other sources of variability, such as fixed oxide charge or charged interface states are expected to take over [59]. Furthermore,

$$\eta \propto \frac{t_{ox}\sigma_{perc}}{A_G}, \tag{1}$$

i.e., in line with basic MOS electrostatics, η will be decreasing with oxide thickness t_{ox} but will increase with decreasing gate area A_G [60]. The technology-dependent enumerator can be determined

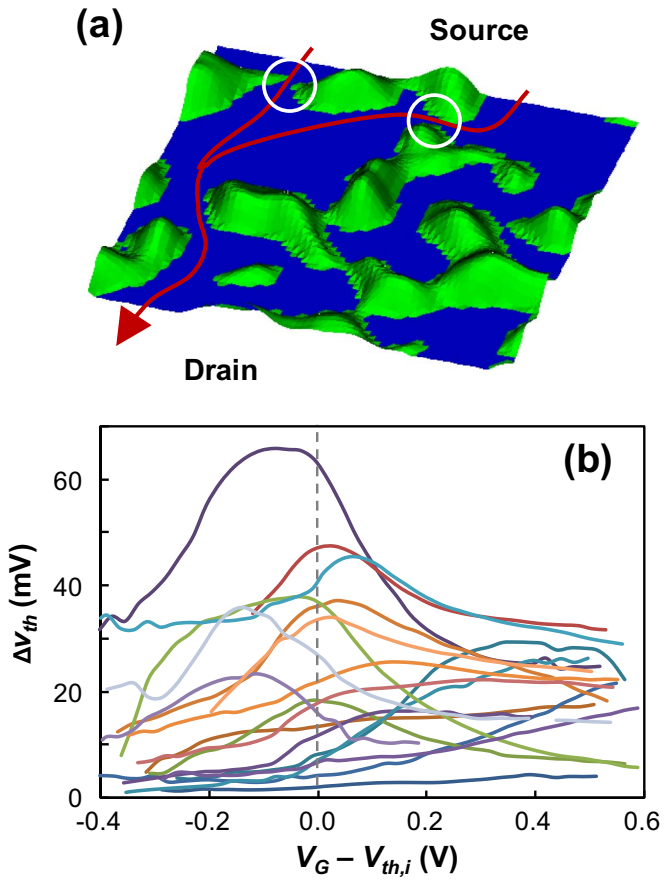

Fig. 10. (a) In small devices, channel conduction near V_{th} proceeds along percolation paths in the channel potential randomized by discrete dopants and other sources of variability. Critical "constriction" spots are demarcated with circles. (b) Widely different $\Delta v_{th}(V_G)$ characteristics are found in nominally identical small-size pFET devices before and after the capture of a single hole, with the shape of the characteristics depending, to a first-order approximation, on the distance between the trap and the nearest constriction spot.

experimentally using e.g. FET arrays or matched pairs [60–62].

In summary, three main quantities, $\langle\tau_c\rangle$, $\langle\tau_e\rangle$, and Δv_{th}, describe the relevant behavior of each gate oxide trap. The capture and emission processes are controlled by tunneling between the carrier reservoir and the trap itself, as well as internal thermal barriers, with both components depending on gate bias V_G, the trap energy level E_T and the depth z_T in the gate oxide. The thermal barriers are further responsible for the temperature dependence of the capture and emission processes. Trap energy levels E_T, depth z_T, and the thermal barriers (determined by parameters S and R) are typically *widely distributed across the trap population*. The individual *instances* of capture and emission events of each trap are stochastic and are *exponentially distributed in time around* $\langle\tau_c\rangle$ and $\langle\tau_e\rangle$, *respectively* [36].

The impact of a charged trap depends, in addition to V_G and z_T, on the lateral position of the trap above the channel. The impact of single traps on the FET threshold voltage Δv_{th} is approximately *exponentially distributed across the trap population*. These distributions are instrumental in propagating the individual trap properties to the device and circuit levels.

4. Implications for devices and circuits

In the previous section we have seen that each trap behaves stochastically around mean values that vary widely from trap to trap. In large-area devices the individual behavior of the many traps averages out and noise measurements are the only way to access the trap

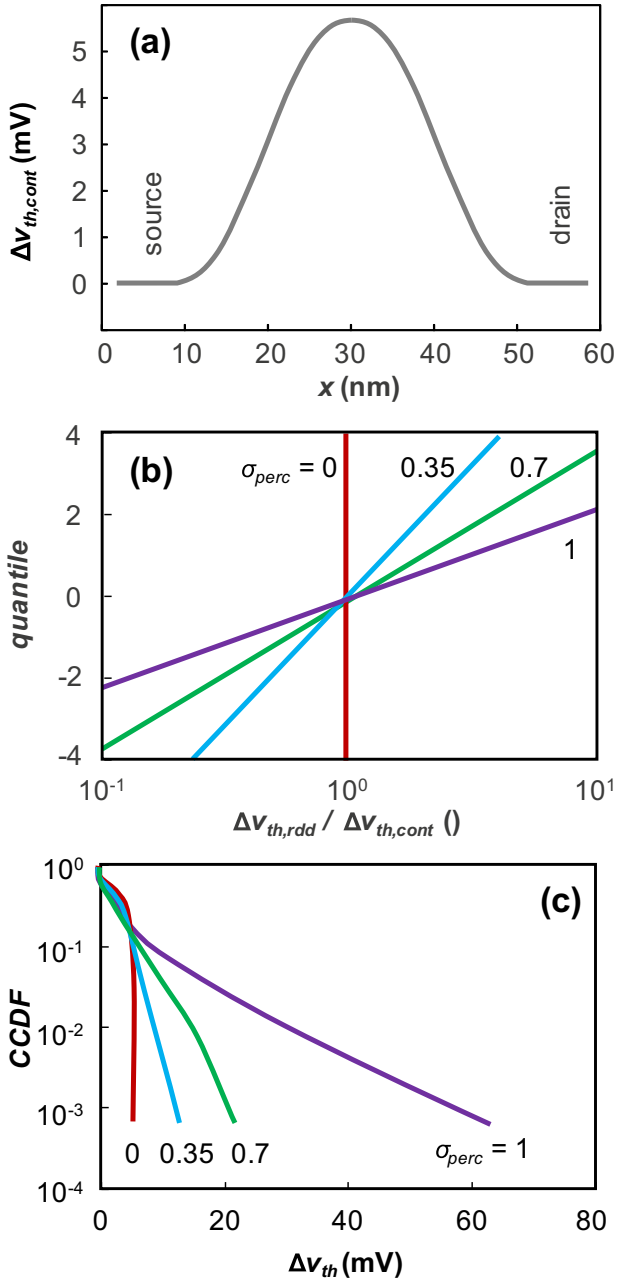

Fig. 11. (a) With realistic *continuous* (cont) FET doping profiles along the channel (direction x), traps close to the S/D junctions contribute less to $\Delta v_{th,cont}$. (b) The ratio of $\Delta v_{th,RDD}$ / $\Delta v_{th,cont}$ constitutes the added impact of Random Dopant Distributions (RDD). The distribution of this ratio can be acceptably described as log-normal, with parameter σ_{perc} (and μ = 0; note: $\sigma_{perc} \equiv 0$ means no additional RDD impact). (c) Complementary CDF plot of a convolution of i) $\Delta v_{th,cont}$ (a) and ii) different amounts of RDD impact, represented by σ_{perc} (b). The tails of the distributions are approximately exponential up to ~ the 99.9th percentile (= 1–10^{-3}).

stochastic properties. As a consequence, large-area devices all degrade approximately identically in standard BTI experiments and the lifetime can be defined as a single value, determined by a parameter, e.g. ΔV_{th}, crossing a predetermined failure criterion, Fig. 12a.

With technology downscaling, the oxide thickness was the first dimension to be reduced to nm scales, resulting in the first *distributed* FET degradation mechanism, TDDB [63]. The gate area of state-of-the-art devices is so small that the gate oxide stack of each device contains just a handful of defects. Because i) each defect behaves stochastically and differently, according to the distributions described above, and ii) the number of defects per device fluctuates, BTI degradation in deeply

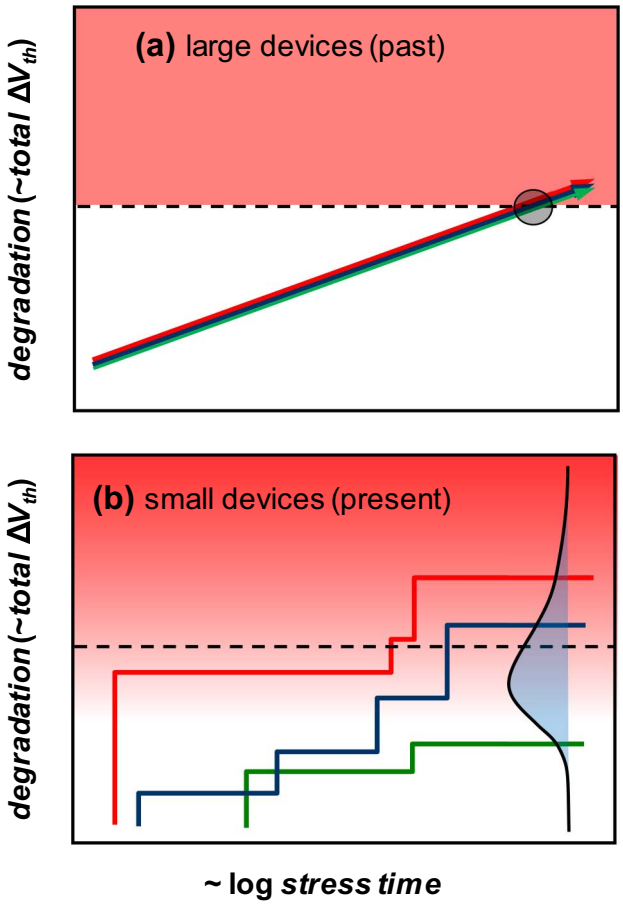

Fig. 12. (a) An illustration of a "conventional" reliability projection. All *large* devices behave identically upon stress (cf. Fig. 6a) and are expected to fail when reaching the projected "hard" degradation criterion ("red zone" above the dashed line independent of the application). (b) Conversely, each *small* device degrades differently, owing to the distributed behavior of a handful of individual traps in it. If the *distribution* of ΔV_{th}'s in all devices at any given time can be inferred, the fraction of devices failing the "hard" criterion can be easily calculated. Circuit-specific failure criteria can be furthermore considered, illustrated here by the graded-red zone. This consideration will allow trading the application performance and reliability margin. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

downscaled devices will be also distributed, i.e. a population of devices will manifest *time-dependent variability*, Fig. 12b. Given the same hard failure criterion as for large-area devices, the lifetime of each small device will be different. Accordingly, a non-zero fraction of devices will exist that surpasses the hard failure criterion and will fail.

With the insights acquired in the previous section we can calculate the distribution of threshold voltage shifts ΔV_{th} in a population of small transistors undergoing identical stress, Fig. 13. We merely need to assume i) the approximately exponential Δv_{th} distribution with mean value η, and ii) Poisson-distributed number of defects per device, with the mean value N. The so-called "defect-centric", or "Exponential-Poisson", distribution of the total device-to-device threshold voltage shifts ΔV_{th} can derived, with its CDF given as

$$H_{\eta,N}(\Delta V_{th}) = \sum_{n=0}^{\infty} \frac{e^{-N}N^n}{n!}\left[1 - \frac{n}{n!}\Gamma\left(n, \frac{\Delta V_{th}}{\eta}\right)\right]. \quad (2)$$

This expression is found to excellently fit the measured ΔV_{th} distributions, Fig. 13 [34,58,60,64,65]. Similar statistics can be observed and similar considerations can be argued for other degradation modes and other FET parameters [66,67].

The advantages of having an analytical description of the distribution are manifold; they include projections to very large and very low

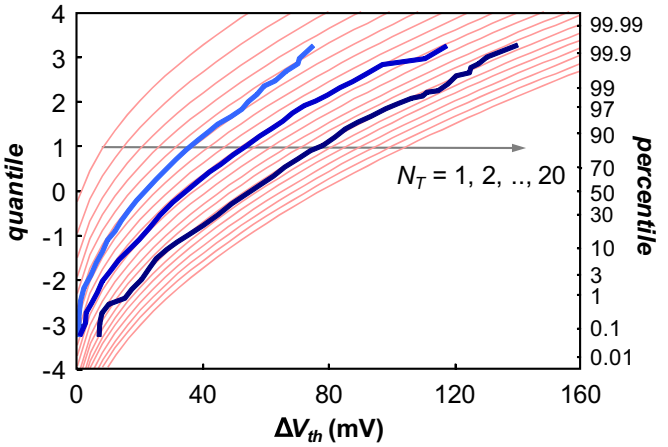

Fig. 13. BTI stress of a population of nominally identical small devices results in wide ΔV_{th} distributions in these devices, ranging from almost 0 to ~100 mV (blue lines: three different levels of stress; data from [13]). All three experimental distributions can be excellently described with the defect-centric distribution (red lines). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

percentiles and analytic forms of the statistical moments. E.g., with the mean degradation being simply

$$\langle \Delta V_{th} \rangle = \eta N, \quad (3)$$

the variance of the ΔV_{th} distribution is conveniently linked with the mean degradation through the mean impact per trap η as

$$\sigma^2_{\Delta V_{th}} = 2\eta \langle \Delta V_{th} \rangle, \quad (4)$$

removing any explicit time dependence [68].

The time-dependent variance, Eq. (4), increases with the mean degradation and may reach values comparable to the as-fabricated, time-zero variance [57,59,69]. This implies that time-dependent variability should be considered during circuit design, in addition to its time-zero counterpart [60,70].

With the defect-centric statistics at hand, the fraction of devices failing a hard ΔV_{th} criterion can be calculated. However, since the failure criterion is in reality circuit-specific, a better approach is to develop methodologies of inserting time-dependent variability into circuit simulations. Such reliability-aware circuit simulations have also the added benefit of allowing to optimize between application performance and reliability margin already during the design phase [42].

Realistic reliability-aware simulations should consider the real workload applied to each device forming the simulated circuit, Fig. 14a [60,65]. Equipped with the physical picture developed above, particularly the trap temporal properties, we can easily calculate the trap occupancy for an arbitrary workload for every device. Several methodology options exist here. We can either assume a large number of traps in the devices and calculate the *mean* degradation $\langle \Delta V_{th} \rangle$ of every device for its specific workload. This in itself may be sufficient information for some more conservative designs. Time-dependent variability can then be calculated based on $\langle \Delta V_{th} \rangle$ using the defect-centric statistics, Fig. 14b. Along this direction, we have developed a methodology to propagate realistic workloads and the statistical distributions through circuit simulation in a deterministic, non-Monte-Carlo manner [71–73].

Optionally, a realistic, i.e., a small number of defects can be instantiated in all deeply scaled devices, allowing to simulate the circuit response at each time step. This approach is useful e. g. for predicting the timing sensitivity of a circuit to RTN. Statistics then needs to be obtained in a Monte-Carlo fashion, through rerunning the simulation multiple times [60,74].

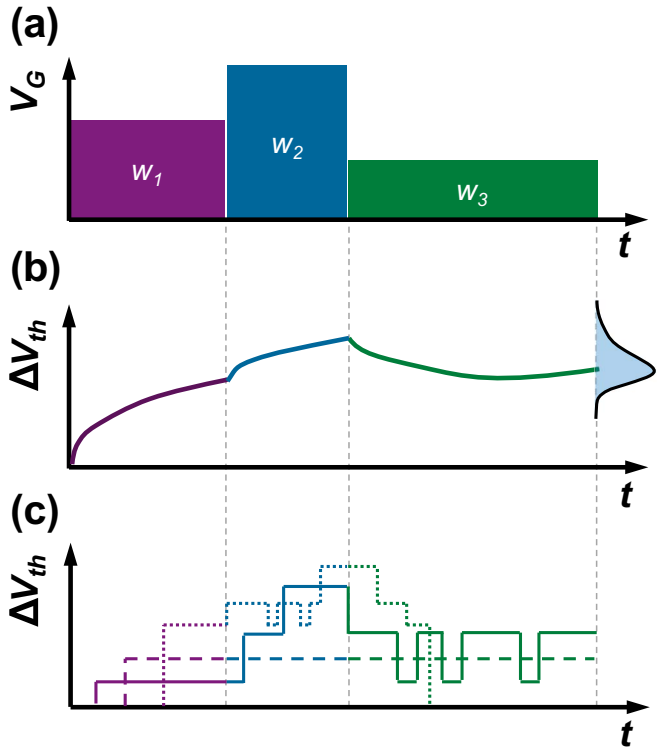

Fig. 14. (a) The workload on each device in a circuit can be approximated by a series of phases w_i with duration t_s, voltage V_G, duty factor *DF*, frequency *f*, and temperature *T*. (In reality, the number of workload phases is typically very large.) (b) For the given workload, the *mean* degradation $\Delta V_{th}(t)$ can be calculated, allowing in turn to obtain the full ΔV_{th} distribution, which can be readily used in circuit analysis. (c) Alternatively, occupancy of *individual* traps and their impact on a circuit can be evaluated as a function of the workload and time.

5. Conclusions

We have reviewed a BTI paradigm based on gate oxide traps and the detailed understanding of their properties. The trap energy levels in the gate dielectric and their misalignment with the channel Fermi level offers the most effective strategy to reduce both PBTI and NBTI in a range of stacks. The trap temporal properties are determined by tunneling between the carrier reservoir and the trap itself, as well as internal thermal barriers of the trap, related to atomic reconfiguration. The electrostatic impact of a trap depends on the gate voltage and its spatial position, randomized by variations in the channel potential. All internal properties of traps are distributed, resulting in distributions of the externally observable parameters $\langle\tau_c\rangle$, $\langle\tau_e\rangle$, and Δv_{th}. These widely distributed parameters in turn result in distributions of the combined effects of ensembles of traps, i.e., in time-dependent variability in devices and circuits.

Acknowledgements

The presented work has been performed within the imec Logic INSITE Partner Program. The work has been in part supported by the European Commission under FP7 project 261868 (MORDRED) and project 619234 (MoRV).

References

[1] B.E. Deal, M. Sklar, A.S. Grove, E.H. Snow, Characteristics of the surface-state charge (Qss) of thermally oxidized silicon, J. Electrochem. Soc. 114 (3) (1967) 266–274.
[2] E.H. Nicollian, J.R. Brews, MOS (Metal Oxide Semiconductor) Physics and Technology, Wiley, 1982.
[3] J. Franco, et al., Characterization of oxide defects in InGaAs MOS gate stacks for high-mobility n-channel MOSFET, Accepted to IEEE Int. Electron Dev. Meeting (IEDM), 2017.
[4] K.O. Jeppson, C.M. Svensson, Negative bias stress of MOS devices at high electric fields and degradation of MNOS devices, J. Appl. Phys. 48 (5) (1977) 2004–2014.
[5] J.H. Stathis, S. Zafar, The negative bias temperature instability in MOS devices: a review, Microelectron. Reliab. 46 (2–4) (2006) 270–286.
[6] E. Cartier, et al., Fundamental aspects of HfO_2-based high-k metal gate stack reliability and implications on t_{inv}-scaling, IEEE Int. Electron Dev. Meeting (IEDM) Tech Dig, 2011.
[7] M.A. Alam, S. Mahapatra, A comprehensive model of PMOS NBTI degradation, Microelectron. Reliab. 45 (1) (2005) 71–81.
[8] T. Grasser, et al., Gate-sided hydrogen release as the origin of "permanent" NBTI degradation: from single defects to lifetimes, IEEE Int. Electron Dev. Meeting (IEDM) Tech Dig, 2015.
[9] R. Gao, et al., Predictive As-grown-Generation (A-G) model for BTI-induced device/circuit level variations in nanoscale technology nodes, IEEE Int. Electron Dev. Meeting (IEDM) Tech Dig, 2016.
[10] M.J. Kirton, M.J. Uren, Noise in solid-state microstructures: a new perspective on individual defects, interface states and low-frequency (1/f) noise, Adv. Phys. 38 (1989) 367–468.
[11] T. Grasser, Stochastic charge trapping in oxides: from random telegraph noise to bias temperature instabilities, Microelectron. Reliab. 52 (2012) 39.
[12] A. Asenov, A.R. Brown, J.H. Davies, S. Kaya, G. Slavcheva, Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs, IEEE Trans. Electron Dev. 50 (9) (2003) 1837–1852.
[13] V. Huard, et al., NBTI degradation: from transistor to SRAM arrays, Proc. Int. Reliab. Phys. Symp. (IRPS), 2008, pp. 289–300.
[14] K. Mistry, et al., A 45 nm logic technology with high-k + metal gate transistors, strained silicon, 9 Cu interconnect layers, 193 nm dry patterning, and 100% Pb-free packaging, IEEE Int. Electron Dev. Meeting (IEDM) Tech Dig, 2007.
[15] B. Kaczer, A. Veloso, M. Aoulaiche, G. Groeseneken, Significant reduction of Positive Bias Temperature Instability in high-k/metal-gate nFETs by incorporation of rare earth metals, Microelectron. Eng. 86 (7–9) (2009) 1894–1896.
[16] M. Aoulaiche, et al., Positive and negative bias temperature instability on La_2O_3 and Al_2O_3 capped high-k MOSFETs, Proc. Int. Reliab. Phys. Symp. (IRPS), 2009.
[17] B. Kaczer, J. Franco, J. Mitard, J. Roussel, A. Veloso, G. Groeseneken, Improvement in NBTI reliability of Si-passivated Ge/high-k/metal-gate pFETs, Microelectron. Eng. 86 (2009) 1582–1584, http://dx.doi.org/10.1016/j.mee.2009.03.061.
[18] J. Franco, et al., SiGe channel technology: superior reliability toward ultrathin EOT devices—part I: NBTI, IEEE Trans. Electron Dev. 60 (1) (2013) 396–404.
[19] R. Degraeve, et al., Trap Spectroscopy by Charge Injection and Sensing (TSCIS): a quantitative electrical technique for studying defects in dielectric stacks, IEEE Int. Electron Dev. Meeting (IEDM) Tech Dig, 2008, pp. 1–4.
[20] G. Sereni, L. Larcher, B. Kaczer, M.I. Popovici, Extraction of the defect distributions in DRAM capacitor using I – V and C – V sensitivity maps, IEEE Electron Dev. Lett. 37 (10) (2016) 1280–1283.
[21] T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, M. Toledano Luque, M. Nelhiebel, The paradigm shift in understanding the bias temperature instability: from reaction–diffusion to switching oxide traps, IEEE Trans. Electron Dev. 58 (11) (2011) 3652–3666.
[22] K. Xiong, J. Robertson, Passivation of oxygen vacancy states in HfO_2 by nitrogen, J. Appl. Phys. 99 (4) (2006) 044105.
[23] D. Liu, J. Robertson, Passivation of oxygen vacancy states and suppression of Fermi pinning in HfO_2 by La addition, Appl. Phys. Lett. 94 (2009) 042904.
[24] S. Sahhaf, et al., Correlation between the V_{th} adjustment of nMOSFETs with HfSiO gate oxide and the energy profile of the bulk trap density, IEEE Electron Dev. Lett. 31 (4) (2010) 272–274.
[25] H. Arimura, et al., Ge nFET with high electron mobility and superior PBTI reliability enabled by monolayer-Si surface passivation and La-induced interface dipole formation, IEEE Int. Electron Dev. Meet. (IEDM) Tech Dig, 2015, pp. 21.6.1–21.6.4.
[26] H. Arimura, et al., Performance and electrostatic improvement by high-pressure anneal on Si-passivated strained Ge pFinFET and gate all around devices with superior NBTI reliability, VLSI Symp. Tech. Dig, 2017, pp. T196–T197.
[27] G. Rzepa, et al., Efficient physical defect model applied to PBTI in high-κ stacks, Proc. Int. Reliab. Phys. Symp. (IRPS), 2017 (pp. XT-11.1–XT-11.6).
[28] B. Linder, et al., Process optimizations for NBTI/PBTI for future replacement metal gate technologies, Proc. Int. Reliab. Phys. Symp. (IRPS), 2016 (pp. 4B-1-1–4B-1-5).
[29] J. Franco, et al., Understanding the suppressed charge trapping in relaxed- and strained-Ge/SiO_2/HfO_2 pMOSFETs and implications for the screening of alternative high-mobility substrate/dielectric CMOS gate stacks, IEEE Int. Electron Dev. Meeting (IEDM) Tech Dig, 2013, pp. 15.2.1–15.2.4.
[30] S. Rangan, N. Mielke, E.C.C. Yeh, Universal recovery behavior of negative bias temperature instability, IEEE Int. Electron Dev. Meeting (IEDM) Tech Dig, 2003, p. 341.
[31] H. Reisinger, O. Blank, W. Heinrigs, A. Mühlhoff, W. Gustin, C. Schlünder, Analysis of NBTI degradation- and recovery-behavior based on ultra fast V_T-measurements, Proc. Int. Reliab. Phys. Symp. (IRPS), 2006, p. 448.
[32] Z. Ji, J.F. Zhang, M.H. Chang, B. Kaczer, G. Groeseneken, An analysis of the NBTI-induced threshold voltage shift evaluated by different techniques, IEEE Trans. Electron Dev. 56 (5) (2009) 1086–1093.
[33] B. Kaczer, T. Grasser, Ph.J. Roussel, J. Martin-Martinez, R. O'Connor, B.J. O'Sullivan, G. Groeseneken, Ubiquitous relaxation in BTI stressing—new evaluation and insights, Proc. Int. Reliab. Phys. Symp. (IRPS), 2008, pp. 20–27.
[34] B. Kaczer, T. Grasser, Ph.J. Roussel, J. Franco, R. Degraeve, L.-A. Ragnarsson, E. Simoen, G. Groeseneken, H. Reisinger, Origin of NBTI variability in deeply scaled pFETs, Proc. Int. Reliab. Phys. Symp. (IRPS), 2010, pp. 26–32.

[35] T. Grasser, H. Reisinger, P.-J. Wagner, F. Schanovsky, W. Goes, B. Kaczer, The time dependent defect spectroscopy (TDDS) for the characterization of the bias temperature instability, Proc. Int. Reliab. Phys. Symp. (IRPS), 2010, pp. 16–25.
[36] T. Grasser, K. Rott, H. Reisinger, M. Waltl, F. Schanovsky, B. Kaczer, NBTI in nanoscale MOSFETs—the ultimate modeling benchmark, IEEE Trans. Electron. Dev. 61 (11) (2014) 3586–3593.
[37] J. Chen, T. Tanamoto, H. Noguchi, Y. Mitani, Further investigations on traps stabilities in random telegraph signal noise and the application to a novel concept physical unclonable function (PUF) with robust reliabilities, Symp. VLSI Technol, 2015.
[38] G. Wirth, R. da Silva, P. Srinivasan, J. Krick, R. Brederlow, Statistical model for MOSFET low-frequency noise under cyclo-stationary conditions, IEEE Int. Electron Dev. Meeting (IEDM) Tech Dig, 2009, p. 715.
[39] M. Toledano-Luque, B. Kaczer, P.J. Roussel, T. Grasser, G.I. Wirth, J. Franco, C. Vrancken, N. Horiguchi, G. Groeseneken, Response of a single trap to AC negative bias temperature stress, Proc. Int. Reliab. Phys. Symp. (IRPS), 2011, pp. 1–8.
[40] M. Toledano-Luque, et al., Temperature and voltage dependences of the capture and emission times of individual traps in high-k dielectrics, Microelectron. Eng. 88 (7) (2011) 1243–1246.
[41] T. Grasser, K. Rott, H. Reisinger, P.-J. Wagner, W. Goes, F. Schanovsky, M. Waltl, M. Toledano-Luque, B. Kaczer, Advanced characterization of oxide traps: the dynamic time-dependent defect spectroscopy, Proc. IEEE Int. Reliab. Phys. Symp. (IRPS), 2013 (pp. 2D.2.1–2D.2.7).
[42] B. Kaczer, C. Chen, P. Weckx, Ph.J. Roussel, M. Toledano-Luque, J. Franco, M. Cho, J. Watt, K. Chanda, G. Groeseneken, T. Grasser, Maximizing reliable performance of advanced CMOS circuits—a case study, Proc. Int. Reliab. Phys. Symp. (IRPS), 2014, p. 2D.4.1.
[43] T. Grasser, et al., On the microscopic structure of hole traps in pMOSFETs, IEEE Int. Electron Dev. Meeting (IEDM) Tech Dig, 2014.
[44] T. Grasser, T. Aichinger, G. Pobegen, H. Reisinger, P.-J. Wagner, J. Franco, M. Nelhiebel, B. Kaczer, The 'permanent' component of NBTI: composition and annealing, Proc. Int. Reliab. Phys. Symp. (IRPS), 2011, pp. 605–613.
[45] K. Huang, A. Rhys, Theory of light absorption and non-radiative transitions in F-centres, Proc. R. Soc. A 204 (1950) 406–423.
[46] C.H. Henry, D.V. Lang, Nonradiative capture and recombination by multiphonon emission in GaAs and GaP, Phys. Rev. B 15 (2) (1977) 989–1016.
[47] W.B. Fowler, J.K. Rudra, M.E. Zvanut, F.J. Feigl, Hysteresis and Franck-Condon relaxation in insulator-semiconductor tunneling, Phys. Rev. B 41 (12) (1990) 8313–8317.
[48] A. Avellán, D. Schroeder, W. Krautschneider, Modeling random telegraph signals in the gate current of metal-oxide-semiconductor field effect transistors after oxide breakdown, J. Appl. Phys. 94 (1) (2003) 703–708.
[49] G. Rzepa, M. Waltl, W. Goes, B. Kaczer, T. Grasser, Microscopic oxide defects causing BTI, RTN, and SILC on high-k FinFETs, Int. Conf. Sim. Semicond. Processes Devices (SISPAD), 2015.
[50] L. Gerrer, J. Ding, S.M. Amoroso, F. Adamu-Lema, R. Hussin, D. Reid, C. Millar, A. Asenov, Modelling RTN and BTI in nanoscale MOSFETs from device to circuit: a review, Microelectron. Reliab. 54 (4) (2014) 682–697.
[51] G. Rzepa, W. Goes, G. Rott, K. Rott, M. Karner, C. Kernstock, B. Kaczer, H. Reisinger, T. Grasser, Physical modeling of NBTI: from individual defects to devices, Int. Conf. Sim. Semicond. Processes Devices (SISPAD), 2014, pp. 81–84.
[52] W. Goes, M. Toledano-Luque, O. Baumgartner, M. Bina, F. Schanovsky, B. Kaczer, T. Grasser, Understanding correlated drain and gate current fluctuations, Proc. of Int. Symp. Phys. Failure and Analysis Integrated Circuits (IPFA), 2013, pp. 51–56.
[53] Bury Erik, R. Degraeve, M. Cho, B. Kaczer, W. Goes, T. Grasser, N. Horiguchi, G. Groeseneken, Study of (correlated) trap sites in SILC, BTI and RTN in SiON and HKMG devices, Proc. of Int. Symp. Phys. Failure and Analysis Integrated Circuits (IPFA), 2013, pp. 250–253.
[54] K. Sonoda, K. Ishikawa, T. Eimori, O. Tsuchiya, Discrete dopant effects on statistical variation of random telegraph signal magnitude, IEEE Trans. Electron Dev. 54 (8) (2007) 1918–1925.
[55] J. Franco, B. Kaczer, M. Toledano-Luque, M.F. Bukhori, Ph.J. Roussel, J. Mitard, L.-Å. Ragnarsson, L. Witters, T. Chiarella, M. Togo, N. Horiguchi, T. Grasser, A. Asenov, G. Groeseneken, Impact of single charged gate oxide defects on the performance and scaling of nanoscaled FETs, Proc. Int. Reliab. Phys. Symp. (IRPS), 2012, p. 5A.4.1.
[56] B. Kaczer, S.M. Amoroso, R. Hussin, A. Asenov, J. Franco, P. Weckx, Ph.J. Roussel, G. Rzepa, T. Grasser, N. Horiguchi, On the distribution of the FET threshold voltage shifts due to individual charged gate oxide defects, IEEE Int. Integrated Reliability Workshop (IIRW) Final Report, 2016, pp. 18–20.
[57] B. Kaczer, et al., Benchmarking time-dependent variability of junctionless nanowire FETs, Proc. Int. Reliab. Phys. Symp. (IRPS), 2017 (pp. 2D-6.1–2D-6.7).
[58] B. Kaczer, Ph.J. Roussel, T. Grasser, G. Groeseneken, Statistics of multiple trapped charges in the gate oxide of deeply scaled MOSFET devices—application to NBTI, IEEE Electron Dev. Lett. 31 (5) (2010) 411–413.
[59] M. Toledano-Luque, B. Kaczer, J. Franco, Ph.J. Roussel, M. Bina, T. Grasser, M. Cho, P. Weckx, G. Groeseneken, Degradation of time dependent variability due to interface state generation, Symp. VLSI Technol. (2013) (pp. T190–T191).
[60] B. Kaczer, et al., The defect-centric perspective of device and circuit reliability—from gate oxide defects to circuits, Solid State Electron. 125 (2016) 52–62.
[61] M. Simicic, A. Subirats, P. Weckx, B. Kaczer, J. Franco, P. Roussel, D. Linten, A. Thean, G. Groeseneken, G. Gielen, Comparative experimental analysis of time-dependent variability using a transistor test array, Proc. Int. Reliab. Phys. Symp. (IRPS), 2016 (pp. XT-10-1–XT-10-6).
[62] B. Kaczer, J. Franco, Ph.J. Roussel, G. Groeseneken, T. Chiarella, N. Horiguchi, Tibor Grasser, Extraction of the random component of time-dependent variability using matched pairs, IEEE Electron Dev. Lett. 36 (4) (2015) 300–302.
[63] R. Degraeve, G. Groeseneken, R. Bellens, M. Depas, H.E. Maes, A consistent model for the thickness dependence of intrinsic breakdown in ultra-thin oxides, IEEE Int. Electron Dev. Meeting (IEDM) Tech. Dig, 1995, p. 863.
[64] C. Prasad, M. Agostinelli, J. Hicks, S. Ramey, C. Auth, K. Mistry, S. Natarajan, P. Packan, I. Post, S. Bodapati, M. Giles, S. Gupta, S. Mudanai, K. Kuhn, Bias temperature instability variation on SiON/Poly, HK/MG and trigate architectures, Proc. IEEE Int. Reliab. Phys. Symp. (IRPS), 2014 (pp. 6A.5.1–6A.5.7).
[65] B. Kaczer, et al., The defect-centric perspective of device and circuit reliability—from individual defects to circuits, 45th European Solid State Device Research Conference (ESSDERC), 2015, pp. 218–225.
[66] B. Kaczer, et al., Origins and implications of increased channel hot carrier variability in nFinFETs, Proc. IEEE Int. Reliab. Phys. Symp. (IRPS), 2015 (pp. 3B.5.1–3B.5.6).
[67] J. Franco, et al., Statistical model of the NBTI-induced threshold voltage, subthreshold swing, and transconductance degradations in advanced p-FinFETs, IEEE Int. Electron Dev. Meeting (IEDM) Tech Dig, 2016.
[69] P. Weckx, B. Kaczer, P. Raghavan, J. Franco, M. Simicic, P.J. Roussel, D. Linten, A. Thean, D. Verkest, F. Catthoor, G. Groeseneken, Characterization and simulation methodology for time-dependent variability in advanced technologies, IEEE Custom Integrated Circuits Conference (CICC), 2015, pp. 1–8.
[68] C. Monzio Compagnoni, R. Gusmeroli, A.S. Spinelli, A. Visconti, Analytical model for the electron-injection statistics during programming of nanoscale NAND flash memories, IEEE Trans. Electron. Dev. 55 (11) (2008) 3192–3199.
[70] P. Weckx, B. Kaczer, M. Toledano-Luque, P. Raghavan, J. Franco, P.J. Roussel, G. Groeseneken, F. Catthoor, Implications of BTI-induced time-dependent statistics on yield estimation of digital circuits, IEEE Trans. Electron. Dev. 61 (3) (2014) 666–673.
[71] P. Weckx, B. Kaczer, H. Kukner, Ph.J. Roussel, P. Raghavan, F. Catthoor, G. Groeseneken, Non-Monte-Carlo methodology for high-sigma simulations of circuits under workload-dependent BTI degradation—application to 6T SRAM, Proc. Int. Reliab. Phys. Symp. (IRPS), 2014 (pp. 5D.2.1–5D.2.6).
[72] D. Rodopoulos, P. Weckx, M. Noltsis, F. Catthoor, D. Soudris, Atomistic pseudo-transient BTI simulation with inherent workload memory, IEEE Trans. Device Mater. Reliab. 14 (2) (2014) 704–714.
[73] D. Stamoulis, et al., Capturing true workload dependency of BTI-induced degradation in CPU components, Proc. Great Lakes Symposium on VLSI (GLSVLSI2016), 2016, pp. 373–376.
[74] B. Kaczer, S. Mahato, V. Valduga de Almeida Camargo, M. Toledano-Luque, Ph.J. Roussel, T. Grasser, F. Catthoor, P. Dobrovolny, P. Zuber, G. Wirth, G. Groeseneken, Atomistic approach to variability of bias-temperature instability in circuit simulations, Proc. IEEE Int. Reliab. Phys. Symp. (IRPS), 2011, pp. XT.3.1–XT.3.5.