Defect-based Methodology for Workload-dependent Circuit Lifetime Projections – Application to SRAM

P. Weckx[1,2], B. Kaczer[2], M. Toledano-Luque[2], T. Grasser[3], Ph. J. Roussel[2], H. Kukner[1,2], P. Raghavan[2], F. Catthoor[1,2], G. Groeseneken[1,2]

[1]Katholieke Universiteit Leuven, ESAT-MICAS, Leuven, Belgium
[2]imec vzw, Leuven, Belgium
[3]TU Wien, Austria
+3216281342, pieter.weckx@imec.be

Abstract— **Despite a number of recent advances made in understanding bias temperature instability (BTI), there is still no simple simulation methodology available which can capture the impact of BTI degradation on deeply scaled transistors, while incorporating the widely distributed defect parameters. We present a physics-based defect-controlled methodology for projecting defect property distributions into circuit lifetime and performance distributions. This methodology allows evaluating the entire population of traps (from fast to slow recoverable and permanent traps), which results in faster simulation and proper extrapolation towards long operating lifetimes.**

Keywords-component; Bias-temperature instability (BTI), time-dependent variability, single-carrier effects, circuit simulations , capture/emission time (CET) maps, SRAM

I. INTRODUCTION

As CMOS scaling continues into the deca-nanometer range, one can literally count the number of gate oxide defects in each FET. Consequently, the stochastic nature of charge capture to and emission from these defects, combined with the substantial relative impact on the device operation result in a *drastic increase in time-dependent variability of degradation mechanisms such as BTI, RTN, and HCI.* The "defect-centric" approach, i.e., understanding time-dependent variability at both the device and circuit level from the perspective of these individual defects is currently being advocated and validated by several groups [1-6]. BTI threshold voltage (V_{th}) shift and recovery can be understood and accurately modeled from the charging and discharging of individual defect traps located in the gate oxide bulk and near the substrate interface. Implementing the behavior of these single defects into circuit-level simulation faces, however, a number of difficulties. On one hand, measuring and collecting the widely distributed defect parameters is limited by experimental concerns (available time × number of samples) (Fig. 1) and the associated extrapolation towards long lifetimes. On the other hand, monitoring and calculating the occupancy of several defects per device throughout circuit simulation poses severe restrictions on the circuit size and the simulation time [3].

In this paper we therefore present a novel statistical simulation methodology circumventing the aforementioned issues without losing the accurate workload dependent behavior of individual defects resulting in *time-dependent variability.*

This work has been partially supported by the Agency for Innovation by Science and Technology in Flanders (IWT)

II. SIMULATION METHODOLOGY

The following section will describe the methodology used to simulate BTI induced V_{th} shift distributions on a transistor level circuit starting from the underlying number of defects and defect property distributions.

A. Analytical description of defects using CET maps

The proposed methodology to evaluate BTI degradation in circuits is based on the analytical modeling using capture/emission time (CET) maps [7,8] describing the probability density function of broadly distributed defect capture and emission times and their correlations acquired at the V_{high} and V_{low} (~V_{th}) digital operating voltages (Fig. 1b). CET maps can be constructed from experimental eMSM data [7-11] limited by the experimental window. Respectively acquiring accurate data for very short or long stress/relaxation times is technically impossible or requires prolonged observation times. Analytical fitting of the CET map, however, allows extrapolating towards very short and long operating lifetimes. Since the capture time τ_c and emission time τ_e are correlated, they are expressed as 2-component vector $\vec{\mathbf{T}}$. Using the energy to time mapping shown in Eq. 1, the CET vectors for the "permanent" ($\vec{\mathbf{T}}_\mathbf{P}$) and "recoverable" ($\vec{\mathbf{T}}_\mathbf{R}$) BTI components can be expressed in terms of effective activation energy vectors $\vec{\mathbf{E}}_{\mathbf{A,P}}$ and $\vec{\mathbf{E}}_{\mathbf{A,R}}$ by

$$\vec{\mathbf{T}}_\mathbf{R} = \tau_{0,R} \exp\left(\vec{\mathbf{E}}_{\mathbf{A,R}} \Big/ \kappa_B T\right) \quad (1\alpha)$$

$$\vec{\mathbf{T}}_\mathbf{P} = \tau_{0,P} \exp\left(\vec{\mathbf{E}}_{\mathbf{A,P}} \Big/ \kappa_B T\right) \quad (1\beta)$$

using the non-radiative multiphonon model for charge exchange [11], where an effective prefactor τ_0 is used, κ_B is the Boltzmann constant and T the temperature. As the effective activation energy can be expressed as a bimodal bivariate normal distribution (Fig. 2), the CET data is accordingly fitted to a bimodal bivariate *log-normal* distribution representing the "permanent" and "recoverable"

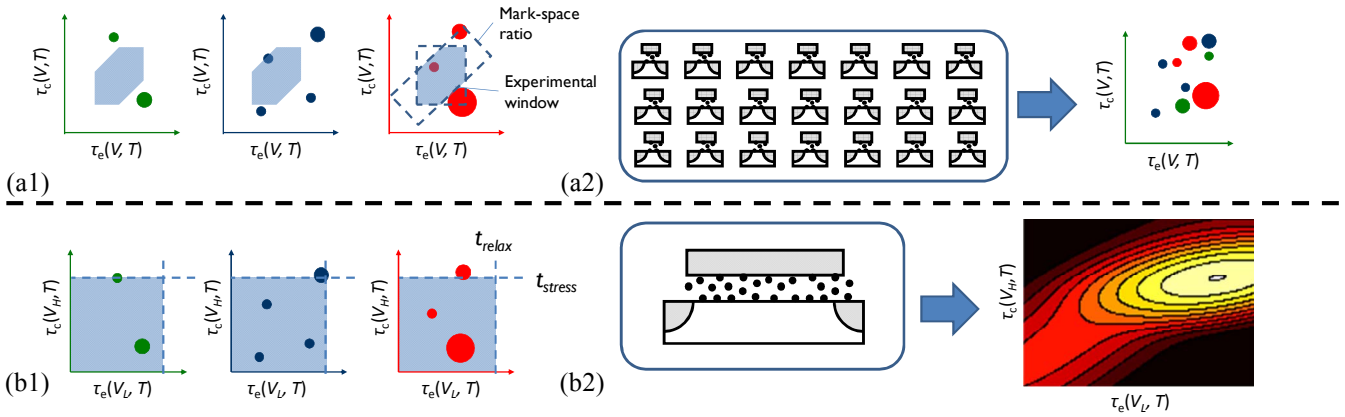

Figure 1. (a1) Typical RTN measurements at fixed voltage can cover only a limited capture and emission space due to the experimental time-window and the mark-space ratio. Depicted are 3 devices characterized by a number of defects (circles), τ_c, τ_e, and the impact on the device when charged (size of the circle). Moreover, (a2) creating a full picture of the τ_e and τ_c space requires elaborate measurements of many small devices. (b1) Alternatively, eMSM covers a wider space and allows measuring a reduced number of large devices (b2), from which part of the map can be constructed [7-11]. Information about the individual defect impact is however lost due to the charge sheet approximation.

BTI components using Eq. 1 [8]. The total (permanent and recoverable) capture and emission time distribution (vector notation $\vec{\boldsymbol{\tau}}$) is hence given by Eq. 2α with the probability density $f(\tau_c, \tau_e)$ in short notation by Eq. 2β.

$$\vec{\boldsymbol{\tau}} \sim A_R \,\text{Log}-\mathcal{N}(\vec{\boldsymbol{\mu}}_{\tau,R}, \vec{\boldsymbol{\Sigma}}_{\tau,R}) + A_P \,\text{Log}-\mathcal{N}(\vec{\boldsymbol{\mu}}_{\tau,P}, \vec{\boldsymbol{\Sigma}}_{\tau,P}) \quad (2\alpha)$$

$$f(\tau_c, \tau_e) = g(\tau_c, \tau_e, A_R, \vec{\boldsymbol{\mu}}_{\tau,R}, \vec{\boldsymbol{\Sigma}}_{\tau,R}, A_P, \vec{\boldsymbol{\mu}}_{\tau,P}, \vec{\boldsymbol{\Sigma}}_{\tau,P}) \quad (2\beta)$$

A_P and A_R are the amplitudes for the permanent respectively recoverable BTI components, $\vec{\mu}_{\tau,P}$ and $\vec{\mu}_{\tau,P}$ are the mean CET vectors describing the log-normal distribution together with their covariance matrices $\vec{\Sigma}_{\tau,P}$ and $\vec{\Sigma}_{\tau,R}$. (Log-$\mathcal{N}$ stands for the lognormal distribution.)

Integrating of the CET map over the entire time domain gives the total defect density n_T (Eq. 3α).

$$n_T = \iint f(\tau_c, \tau_e) d\tau_c d\tau_e \quad (3\alpha)$$

$$N_T = WLn_T \quad (3\beta)$$

This density can be rescaled using Eq. 3β to the simulated device dimensions, giving the mean number of available traps in each device N_T.

B. Trap capture probability

Previously, it has been shown that CET maps can accurately describe stress and recovery patterns for DC, AC and DF stressing by integrating over the part of the CET map that is active under stress [8,9,12]. Essentially for each trap characterized by $\vec{\boldsymbol{\tau}}$ on the CET map, the occupancy probability P_{occ}, depending on the applied stress waveform, is evaluated and the trap subsequentially contributes to the overall V_{th} shift. Consequently, only a fraction of the trap population will be active, i.e., taking part in the degradation of the specified device. The occupancy probability for AC and DF stressing between two digital voltages (high (V_H) and low (V_L)) can be readily calculated using Eq. 4 for a given frequency f, duty factor DF and stress time t_{stress}. $P_{occ,H}$ is the probability for a trap to be occupied after the high voltage V_H is applied. Arbitrary (non-periodic) waveforms require piecewise evaluation of each V_L relax and V_H stress period as shown in Fig. 3 [13,14].

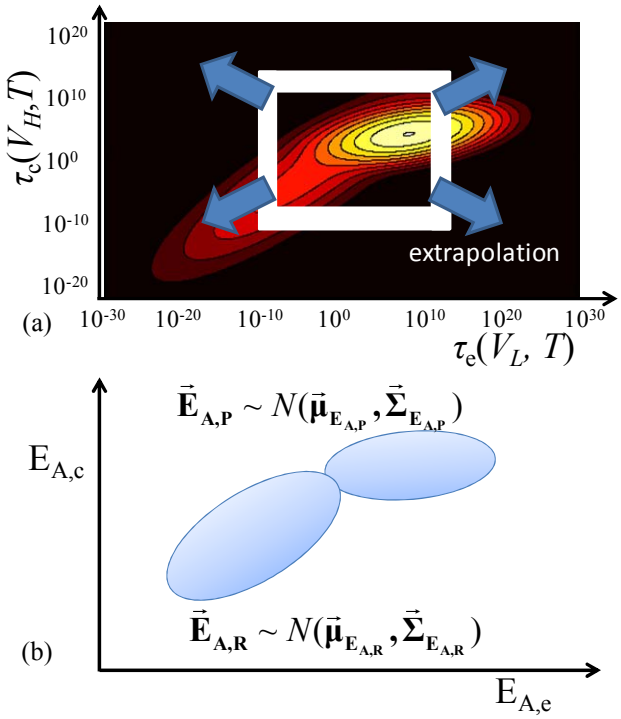

Figure 2. (a) Analytical fitting of the CET data to a bimodal bivariate log-normal distribution, (describing the "permanent" and "recoverable" BTI components) corresponds to (b) fitting the effective activation energy data to a bimodal bivariate normal distribution using Eq. 1. [8]

The occupancy probability for a AC stress:

$$P_{occ,H} = \frac{1-e^{-\frac{DF}{f\tau_c}}}{1-e^{-\frac{1}{f}\left(\frac{DF}{\tau_c}+\frac{1-DF}{\tau_e}\right)}}\left(1-e^{-t_{stress}\left(\frac{DF}{\tau_c}+\frac{1-DF}{\tau_e}\right)}\right) \quad (4)$$

as function of the entire τ_e and τ_c domain is graphically shown as a probability map in Fig. 4b. By multiplying the original CET map with this occupancy probability map a resulting *CET-active map* is constructed, describing the distribution of the active traps after the corresponding stress waveform (Fig.4c).

By integrating the CET-active map over the entire time domain and rescaling using

$$\rho = \frac{\iint f(\tau_c,\tau_e)P_{occ,H}(\tau_c,\tau_e,DF,t_{stress},f)d\tau_c d\tau_e}{\iint f(\tau_c,\tau_e)d\tau_c d\tau_e} \quad (5)$$

one can obtain the ratio of traps ρ which will be occupied as a result of the applied stressing waveform.

The mean number of occupied traps $N_{T,occ}$ for a specific device is then calculated from the mean number of available traps in that device N_T (which is proportional to the device area (Eq. 3α)), together with the ratio of occupied traps ρ (dependent on the applied waveform), using Eq. 6.

$$N_{T,occ} = N_T\rho \quad (6)$$

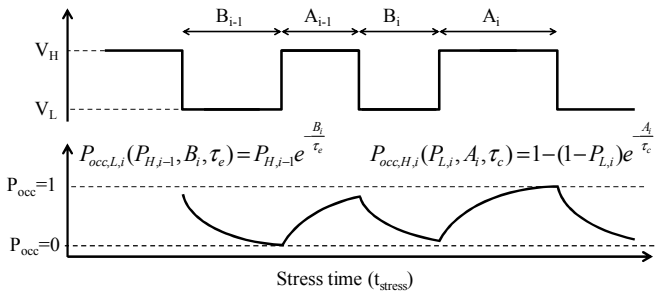

Figure 3. Piecewise evaluation of the occupancy probability after each V_H or V_L digital voltage period is required to calculate the total occupancy probability at a certain time. The occupancy probability at a V_H or V_L period is calculated using the duration of that period and the previous period's trap occupancy probability.

C. *Threshold voltage shift distribution*

The total ΔV_{th} Cumulative Distribution Function (CDF) for a given mean number of traps $N_{T,occ}$ can subsequently be obtained using Eq. 7 [15] where η is the average impact per defect (~1/device area) which can be extracted from experiments [16] and N_T, the mean number of occupied traps

$$H_{\eta,N_T}(\Delta V_{th},\eta) = \sum_{n=0}^{\infty}\frac{e^{-N_T}N_T^n}{n!}\left[1-\frac{n}{n!}\Gamma(n,\Delta V_{th}/\eta)\right] \quad (7)$$

As a result the time constants are eliminated from the threshold voltage distribution calculation which now only depends on the mean number of occupied traps activated by the workload. Figure 5 summarizes the simulation methodology for using a CET map together with AC or DF workloads combined with the transistor dimensions to obtain workload dependent BTI induced V_{th} shifts. (The same methodology also applies for arbitrary waveforms which will only have a different probability map compared to a periodical stress waveform).

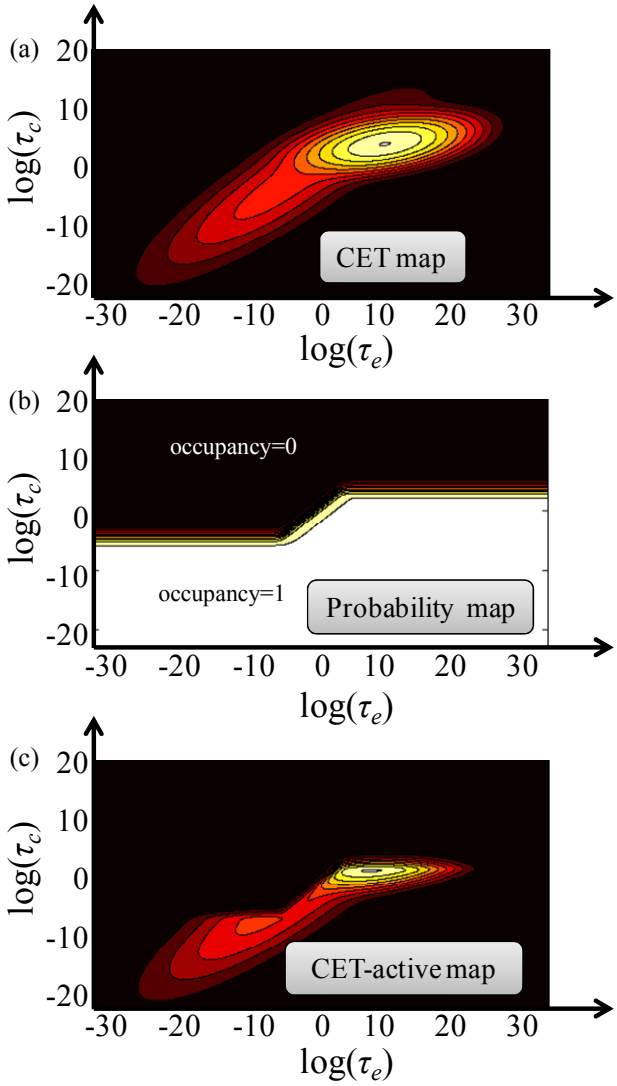

Figure 4. (a) The 2-component CET map, fitted on experimantally extracted data from large devices on the corresponding technology. (b) Occupancy probability map after an AC stress as function of τ_e and τ_c [12,13]. (c) The CET-active map is created by multiplying the occupancy probability map with the analytical CET map.

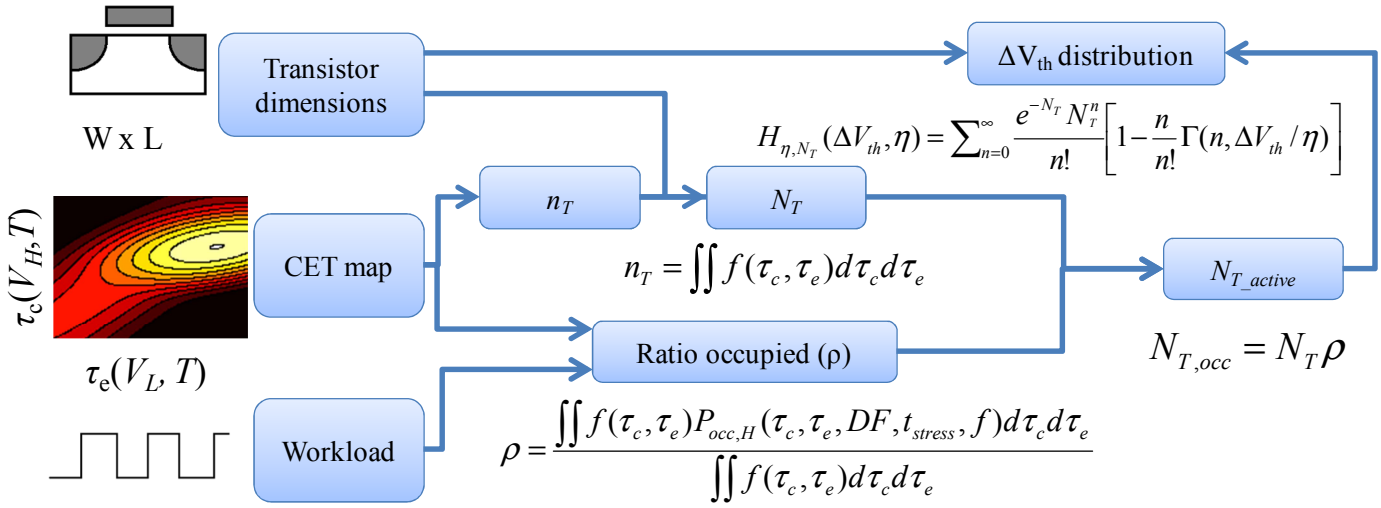

Figure 5. Simulation methodology for degradation using a CET map and arbitrary or AC workloads. By numerical integration of the CET-active map (Eq. 5), the ratio of occupied traps can be calculated, from which, taking into account the device dimensions, the mean number N_T (= n_t × device area) of captured traps is obtained (Eq. 3). Taking N_T together with η, the average impact per defect (~1/device area) which can be extracted from experiments [16], workload induced ΔV_{th} distributions are obtained (Eq. 6) [15].

III. Application to SRAM

The Static Random Access Memory (SRAM) cell is currently present in almost every integrated CMOS product. Since the SRAM cell is usually also the most scaled device of the System-On-Chip (SoC), its reliability is of great importance to evaluate future technology scaling. We expect that the proposed degradation simulation framework should prove useful for quick investigation of the reliability of CMOS circuits. The following section will describe the application of this methodology to the SRAM cell.

A. Experimental setup

Our approach is demonstrated on the case of an SRAM cell depicted in Fig 6a. As the Figure Of Merit (FOM) under survey, SNM is taken from the butterfly plot [Fig. 6b][17] during read mode, i.e. when the Word Line (WL) and Bit Line (BL, $\overline{\text{BL}}$) voltages equal the supply voltage V_{DD}. Depending on their actual workload, the p- and n-channel FETs of the SRAM cell are each linked to a different CET-active map. Only pFET CET maps (corresponding to NBTI) were used for this demonstration. Using the proposed methodology shown in Fig. 5, ΔV_{th} distributions for each pFET can be quickly obtained for a variety of workloads. Time-zero V_{th0} distribution can be subsequently added on top of the BTI induced ΔV_{th} to acquire the total V_{th} distribution.

B. Simulation approach

After acquiring V_{th} distributions for each of the circuit's transistors, a Monte Carlo (MC) sampling is performed where a SPICE level simulator is used to calculate the SNM of the SRAM cell. Different technology nodes are tested using the Predictive Technology Model (PTM) cards for low-power applications [18]. Compared to MC sampling of the CET map combined with calculating the occupancy of several defects per device throughout the circuit simulation described in [3], a significant speed-up is obtained as shown in Fig. 7b. Moreover, as a way to further speed up simulation when a large number of samples is required, response surfaces (i.e. LUTs [Fig. 7a]) are created for the important circuit metrics. Creating these response surfaces does however take an initial investment of simulation time (setup-time) depending on the size of the look-up table. For very large sample sizes however (e.g. Giga samples), the original simulation time becomes much larger than the response surface's setup-time and several orders of magnitude in speed-up can be achieved. This significant reduction in CPU does not come at the expense of accuracy. Shown in Fig. 8 are the V_{th} distributions calculated using the introduced CET methodology (Fig.5) compared with calculating the occupancy of several defects per device in a Monte Carlo circuit simulation [3].

C. Results

Fig. 9 shows the SRAM SNM distributions on a probit plot for different technologies as a function of stress time t_{stress} for an AC workload with Duty Factor (DF)=0.5 and frequency (f)=10^7. Figs. 10a and 10b depict the median SNM $\mu_{1/2}$ degradation and specific spread as a function of t_{stress}. As the metric to show the specific spread the distance between the -1 sigma probability quantile ($q_{-1\sigma}$) and the median $\mu_{1/2}$ is normalized to the median. Increased degradation is observed for more deeply scaled technologies. Figs. 10c and 10d show again the median degradation and specific spread as a function of the workload DF. Figs. 11a and 11b show the median SNM degradation and specific spread as a function of t_{stress} for a 45nm technology taking a 50mV time-zero variation into account.

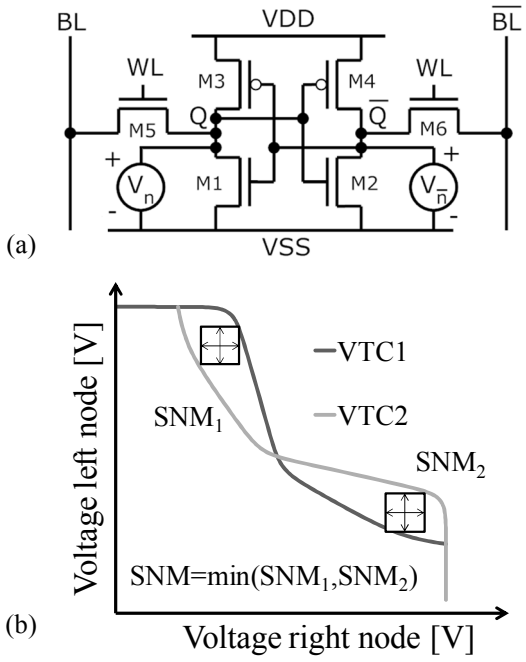

Figure 6. (a) The 6T-SRAM cell used for demonstration. (b) SNM margins are obtained by analyzing the butterfly curve during read mode, where the word and bit lines are set at the V_{DD} which is scaled according to the technology node used.

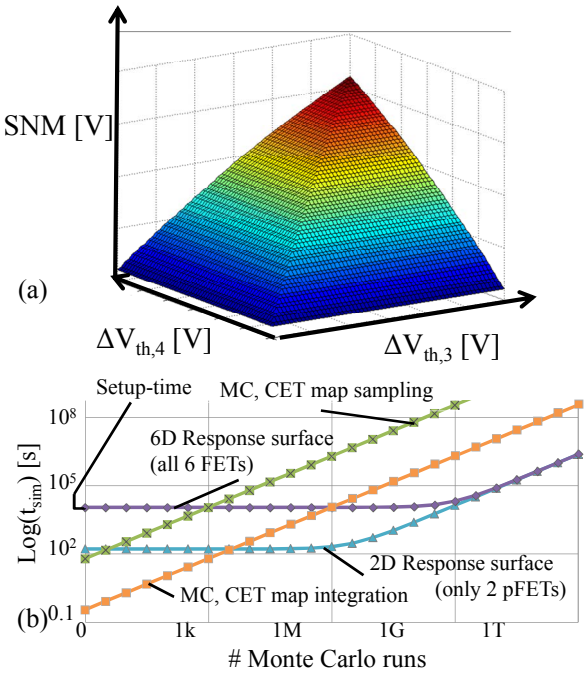

Figure 7. (a) A SNM response surfaces as function of the pFET's V_{th} (b) compared to standard Monte Carlo (MC) sampling of the capture and emission times out of the CET map [3], response surfaces provide a significant speed-up. For very large sample sizes (e.g. Giga samples), several orders of magnitude in speed-up can be achieved, once the setup-time investment returned.

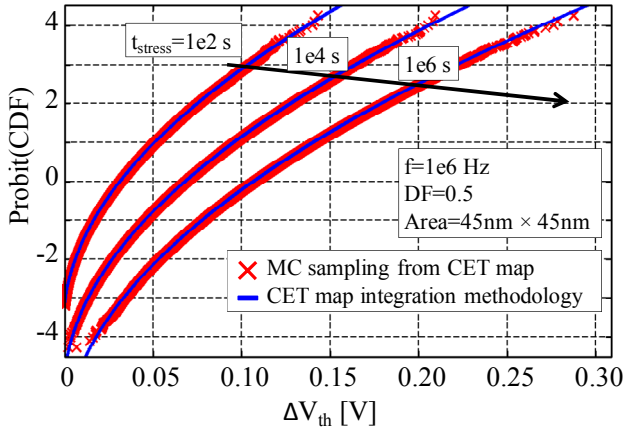

Figure 8. Comparison of the ΔV_{th} cumulative density function calculated via Monte Carlo (MC) sampling of the CET map (red markers) [3] and the integration of the CET-active map (blue lines) for different stress times t_{stress}=10^2 s, 10^4 s, 10^6 s.

D. Discussion

From Fig. 9 it is apparent that scaling down the SRAM cell results in an increased BTI induced degradation of the SNM which in turn results in a lower reliability of large SRAM arrays. This increased degradation can be attributed to two inherent factors of the defect based BTI mechanism. Firstly η (the average impact per defect) is increased due to the ~1/device area dependency [16] and secondly the average number of defects (Eq. 3β) reduces giving rise to an increased spread on the V_{th} shift [15]. With the decreasing supply voltage V_{DD} for each scaled technology node, the mean relative impact of BTI on SNM of the SRAM cell increases as well. This is again demonstrated in Figs 10a and 10b where the median degradation and specific spread of SNM is shown. An increased degradation rate is observed between 10^2 s and 10^5 s as the 'permanent' BTI component is contributing more and more to the overall V_{th} shift of the pFET devices. Increased degradation is additionally observed for DF balanced SRAM cells (i.e., DF ~ 0.5) compared to an unequally stressed cell as shown in Fig. 10c and Fig. 10d.

Taking time zero variability into account, as shown in Fig. 11a, it seems that the initial distribution of SNM is not changed that much by the BTI component. However, looking at Fig 11b we can account for a 10% median degradation of the SNM due to BTI and an increased specific spread as well. Further investigation will be needed to assess what will be the relative impact of BTI on the SRAM stability compared to time-zero variability at higher quantiles.

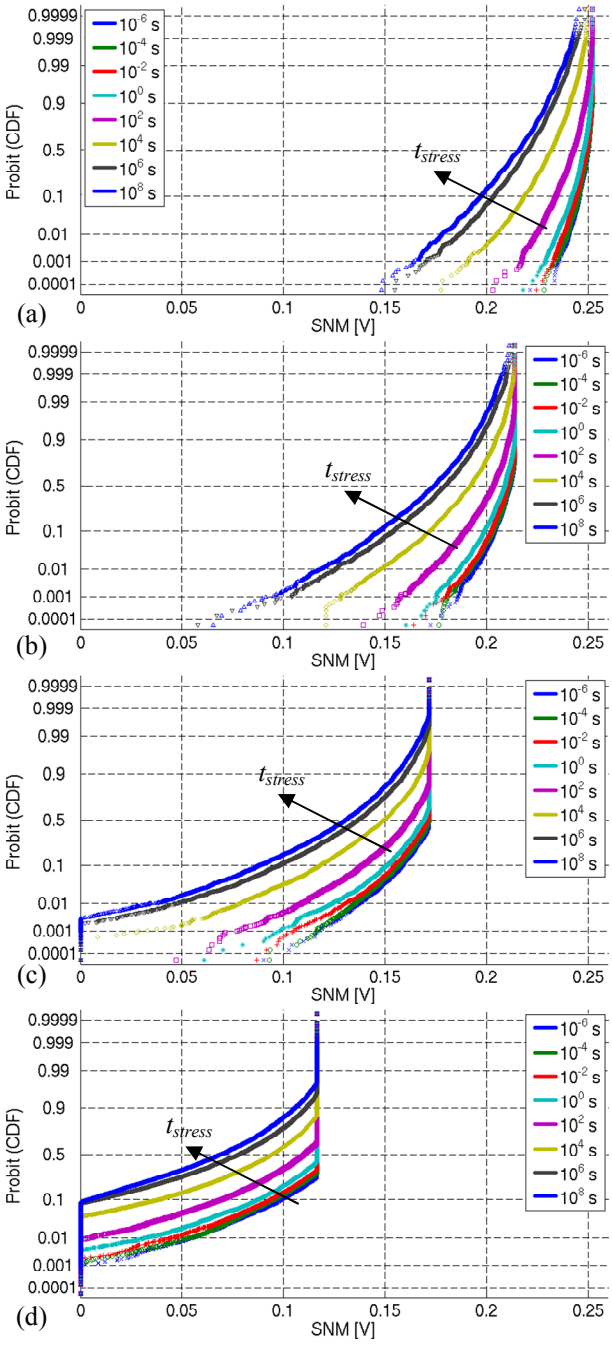

Figure 9. Probit plots of the SRAM SNM distributions for an AC workload with Duty Factor (DF)=0.5 and frequency (f)=10^7 Hz obtained by MC simulations. Plotted are different stress times (t_{stress}) for technologies (a) 45 nm, (b) 32 nm, (c) 22 nm and (d) 16 nm. Time-zero variability is not taken into account.

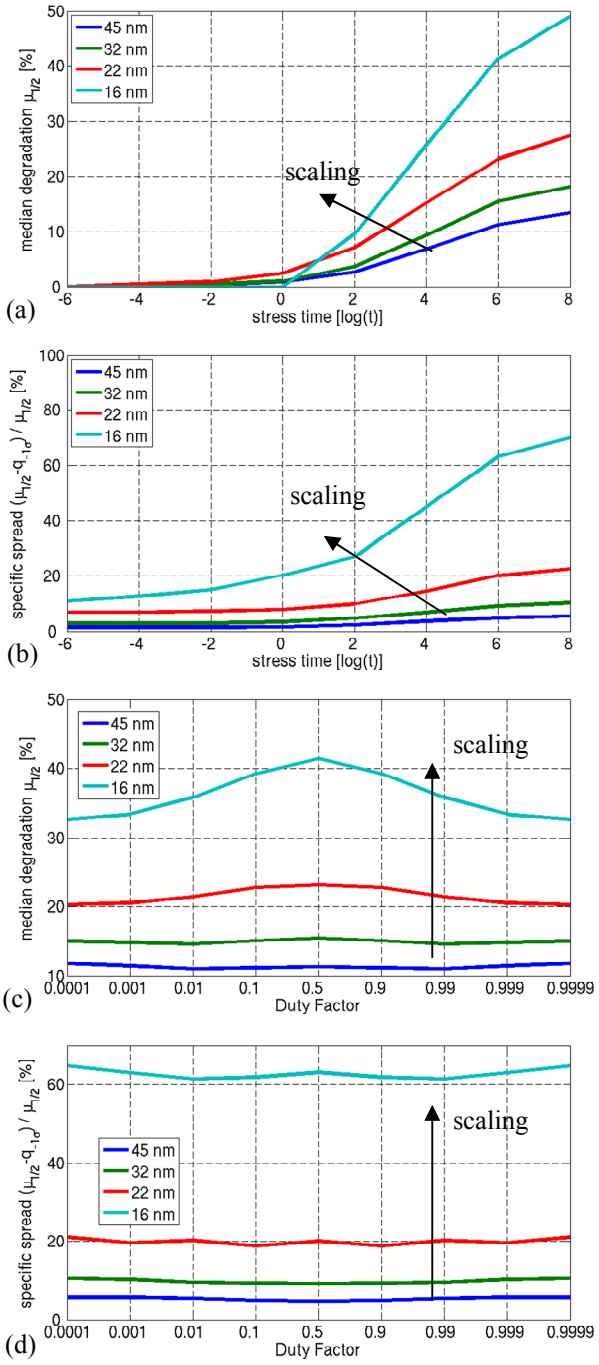

Figure 10. (a) SNM degradation plots for the median degradation and (b) specific spread as function of the t_{stress} for an AC workload with *DF*=0.5 and *f*=10^7 Hz. Increased median degradation and spread are observed for deeply scaled technologies. (c,d) The median degradation and specific spread as function of *DF*, for t_{stress}=10^6 s and *f*=10^7 Hz. Increased DF sensitivity is observed for deeply scaled technologies.

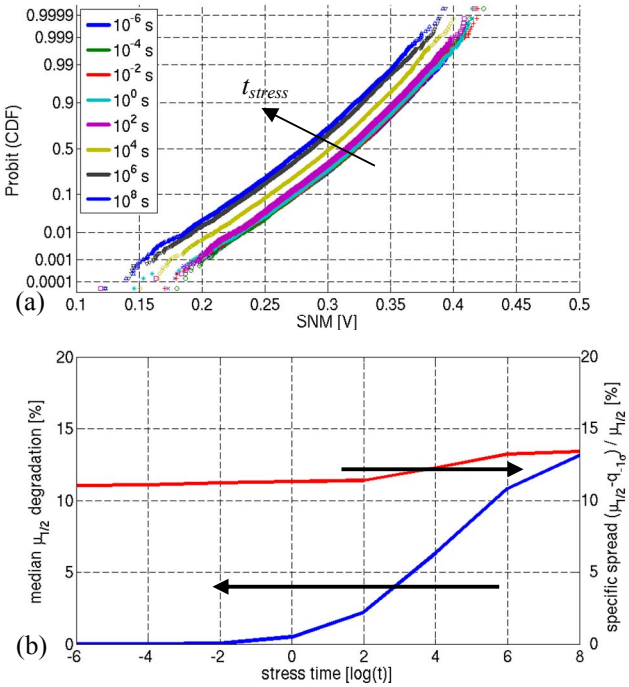

Figure 11. (a) Probit plots of the SRAM SNM distributions for an AC workload with Duty Factor (DF)=0.5 and frequency (f)=10^7 Hz, taking 50mV time zero variability into account, for a 45nm PTM technology. (b) The median degradation and specific spread as a function of the t_{stress} for an AC workload with *DF*=0.5 and *f*=10^7 Hz.

IV. CONCLUSIONS

We have demonstrated a physics-based defect-centric methodology for projecting defect property distributions into circuit lifetime and performance metric distributions. BTI induced degradation is widely understood in a unified manner as the cumulative effect of the capture and emission of individual defect traps. CET map extraction allows evaluating the entire population of traps (from fast traps to slow recoverable and permanent traps), which results in faster characterization, faster simulation and proper extrapolation towards long operating lifetimes. Results have been shown for an SRAM cell where the workload dependent behavior is well established, which is crucial for evaluating the true yield and FOM distribution.

REFERENCES

[1] A. Asenov, S. Roy, R.A. Brown, G. Roy, C. Alexander, C. Riddet, C. Millar, B. Cheng, A. Martinez, N. Seoane, D. Reid, M.F. Bukhori, X. Wang, U. Kovac, "Advanced simulation of statistical variability and reliability in nano CMOS transistors," *IEDM*, 2008, pp. 15-17

[2] A. Islam, A. Muhammad, "Analyzing the distribution of threshold voltage degradation in nanoscale transistors by using reaction-diffusion and percolation theory," *Journal of Computational Electronics,* 2011, pp 341-351

[3] B. Kaczer, S. Mahato, V.V. de Almeida Camargo, M. Toledano-Luque, P.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," *IRPS*, 2011, pp. 10-14

[4] V. Huard, N. Ruiz, F. Cacho, E. Pion, "A bottom-up approach for System-On-Chip reliability," *Microelectronics Reliability*, 2011, pp. 1425–1439

[5] J. Fang, S.S. Sapatnekar, "Understanding the impact of transistor-level BTI variability," *IRPS*, 2012, pp. 15-19

[6] K.V. Aadithya, A. Demir, S. Venugopalan, J. Roychowdhury, "SAMURAI: An accurate method for modelling and simulating non-stationary Random Telegraph Noise in SRAMs," *DATE*, 2011 , pp.1-6

[7] H. Reisinger, T. Grasser, W. Gustin, C. Schlünder, "The statistical analysis of individual defects constituting NBTI and its implications for modeling DC- and AC-stress," *IRPS*, 2010, pp. 7-15

[8] T. Grasser, P. Wagner, H. Reisinger, T. Aichinger, G. Pobegen, M. Nelhiebel, B. Kaczer, "Analytic modeling of the bias temperature instability using capture/emission time maps," *IEDM*, 2011, pp. 1-4

[9] K. Zhao, J.H. Stathis, B.P. Linder, E. Cartier, A. Kerber, "PBTI under dynamic stress: From a single defect point of view," *IRPS*, 2011, pp. 1-9

[10] T. Nagumo, K. Takeuchi, T. Hase, Y. Hayashi, "Statistical characterization of trap position, energy, amplitude and time constants by RTN measurement of multiple individual traps," *IEDM,* 2010, pp. 1-4

[11] T. Grasser, H. Reisinger, P. Wagner, F. Schanovsky, W. Goes, B. Kaczer, "The time dependent defect spectroscopy (TDDS) for the characterization of the bias temperature instability," *IRPS*, 2010, pp.16-25

[12] H. Reisinger, T. Grasser, K. Ermisch, H. Nielen, W. Gustin, C. Schlunder, "Understanding and modeling AC BTI," *IRPS*, 2011, pp.1-8,

[13] J. Martin-Martinez, B. Kaczer, M. Toledano-Luque, R. Rodriguez, M. Nafria, X. Aymerich, G. Groeseneken, "Probabilistic defect occupancy model for NBTI," *IRPS,* pp. 1-6

[14] 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," *IRPS,* 2011, pp. 1-8

[15] B. Kaczer, T. Grasser, P.J. Roussel, J. Franco, R. Degraeve, L. Ragnarsson, E. Simoen, G. Groeseneken, H. Reisinger, "Origin of NBTI variability in deeply scaled pFETs," *IRPS*, 2010, pp. 26-32

[16] J. Franco, B. Kaczer, M. Toledano-Luque, P.J. Roussel, J. Mitard, L. Ragnarsson, L. Witters, T. Chiarella, M. Togo, N. Horiguchi, G. Groeseneken, M.F. Bukhori, T. Grasser, A. Asenov, "Impact of single charged gate oxide defects on the performance and scaling of nanoscaled FETs," *IRPS,* 2012, pp. 1-6

[17] E. Seevinck, F.J. List, J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE Journal of Solid-State Circuits*, 1987, pp. 748- 754

[18] http://ptm.asu.edu/