This is the accepted version of the article:

This version is available at https://ddd.uab.cat/record/249165
under the terms of the (C) BY-NC-ND license
A Smart Noise- and RTN-Removal Method for Parameter Extraction of CMOS Aging Compact Models

Javier Diaz-Fortuny1*, Javier Martin-Martinez1, Rosana Rodriguez1, Rafael Castro-Lopez2, Elisenda Roca2, Francisco V. Fernandez2 and Montserrat Nafria1

1Electronic Engineering Department, (REDEC) group. Universitat Autònoma de Barcelona (UAB), Barcelona 08193, (Spain).
2Instituto de Microelectrónica de Sevilla, IMSE-CNM, CSIC and Universidad de Sevilla, (Spain).

*corresponding author: javier.diaz@uab.es

Abstract—In modern nanometer-scale CMOS technologies, time-zero (TZV) and time-dependent variability (TDV) effects, e.g., coming from aging mechanisms like Bias Temperature Instability (BTI), Hot Carrier Injection (HCI) or Random Telegraph Noise (RTN), have re-emerged as a serious threat affecting the performance of analog and digital integrated circuits. Variability induced by the aging phenomena can lead circuits to a progressive malfunction or failure. In order to understand the effects of the mentioned variability sources, a precise statistical characterization and modeling of these effects should be done. Typically, transistor TDV characterization entails long, and typically prohibitive, testing times, and huge amounts of data, which are complex to post-process. This work presents a new method to statistically characterize the emission times and threshold voltage shifts (ΔVT) related to oxide defects in nanometer CMOS transistors during aging tests. The method identifies the VT drops associated to oxide trap emissions during BTI and HCI aging recovery traces while removing RTN and background noise contributions, to avoid artifacts during data analysis.

Keywords—CMOS; BTI; HCI; parameters; extraction; method; RTN; defects; aging;

I. INTRODUCTION

With nowadays CMOS technology downscaling, BTI [1][2] and HCI [3][4] aging effects, as well as RTN [5][6] transient effects, have re-emerged as important time-dependent variability (TDV) phenomena that must be taken into account in the design of digital and analog integrated circuits (ICs). During circuit operation, the variability effects related to the trapping/detrapping in/from oxide defects could result in circuit malfunction due to the shift of some transistor parameters, such as the threshold voltage (VT) [7]. Thus, it is critical for IC designers to take into account TDV effects to implement reliability-aware circuits [8]. To this end, appropriate TDV compact models, like the Probabilistic Defect Occupancy (PDO) model [3], are essential. These models need to account for and clearly distinguish the ‘slow’ defects, responsible for aging-induced degradation, from the ‘fast’ defects causing the RTN transient variations. As shown in this paper, isolating the impact of the different types of defects within the experimental data is not straightforward.

Another important aspect is that an accurate extraction method of the defect parameters requires statistical characterization of transistors under accelerated stress conditions [9]. Conventional BTI/HCI aging characterization techniques are based in the application of serialized stress-measurement (SM) sequences to one or few transistors simultaneously by using probe stations. The SM characterization technique consists in an stress phase, which accelerates the device degradation through the application of overvoltage to the device terminals, and a measurement phase, where the effects of the induced aging can be measured in reasonable testing times [10]. However, to get sufficient statistical data, a massive aging test, with hundreds of transistors subjected to the same characterization pattern, must be performed. Doing so serially (one device at a time) may, however, take several months. In our work, we use our previously designed 65-nm array-based IC ENDURANCE chip [11] in combination with a custom-designed characterization setup [12] that allows to execute BTI/HCI aging tests over hundreds of CMOS transistors with a stress parallelization technique that significantly reduces the total aging test time.

The impact of aging can be modeled through the analysis of the emission time (τ) and corresponding impact on the variation of the threshold voltage (η) of each defect in the transistor. This analysis is carried out during the measurement phases, after the application of overvoltage stress, where a ‘recovery’ of the device threshold voltage VT can be observed [13]. Fig. 1 shows experimental recovery traces attained after the execution of a massive BTI test, where charge detrapping from oxide defects can be clearly seen during the recovery time as positive abrupt current jumps (i.e., VT decrease). Note also

![Fig. 1. Several recovery traces from a massive BTI test showing discrete current increments.](image-url)
in Fig. 2 that the simultaneous presence of RTN in the traces could mask or significantly increase the current increments, so it is critical to distinguish between those current levels that are linked to fast transients and slow defect discharges.

In this work, we introduce a novel and smart method of defect parameters extraction that identifies the defect-related $\tau_e$ and $\eta$ values from a large number of experimental transistor recovery traces, where fast defect capture/emissions (i.e., RTN) and background noise are present. To model the effect of aging without noise-masking issues, the defect parameters extraction method removes the RTN and background noise from the measured BTI/HCl recovery traces and evaluates the $\{\tau_e, \eta\}$ tuple of each slow defect discharge found.

II. DESCRIPTION OF THE METHOD

The defect parameters extraction method follows the flow diagram shown in Fig. 3. For each device involved in the aging test, the method treats each recovery trace individually. First, it converts the $I_{DS}$ vs. time traces measured after the stress period into the equivalent $\Delta V_{th}$ vs. time trace (the difference between the instantaneous threshold voltage and that of the fresh device). By means of the Weighted Time Lag Plot (WTLP) method [5][6], all $\Delta V_{th}$ levels of the trace are obtained. In order to remove the background noise from the recovery traces, each experimental $\Delta V_{th}$ sample is assigned to one of the $\Delta V_{th}$ levels previously found with the WTLP method, quantizing the trace and removing the background noise. Then, all RTN fast defect transitions (if any) are removed from the trace through the analysis of each transition between the $i$-th and the $(i+1)$-th recovery sample. The last step consists in the extraction of $\tau_e$ and $\eta$ parameters associated to each of the identified defect discharge. This 5-step procedure is described in detail in the following.

1) $I_{DS}$ conversion to $\Delta V_{th}$. For each tested device, the $I_{DS}$ recovery traces measured as a function of time during the aging test is converted into the equivalent $\Delta V_{th}$ vs. time trace by means of the pre-stress $I_{DS}$-$V_{th}$ curve, i.e., the $I_{DS}$-$V_{th}$ curve of the fresh device [14]. For instance, Fig. 4 shows the results of the conversion from a measured $I_{DS}$ trace, like the ones represented in Fig. 1, to its equivalent $\Delta V_{th}$ waveform, where 3 charge emissions (denoted with green arrows) can be clearly located during the measurement time window of 100s.

Moreover, as can be observed in the recovery trace, fast transitions associated to RTN are present during the entire characterization experiment. The inset of Fig. 4 exposes one of those fast defect transitions between two separated $\Delta V_{th}$ levels superimposed to a temporarily constant $\Delta V_{th}$ level. Fig. 4(b) reveals fast capture/emission transitions, switching between 4 different $\Delta V_{th}$ levels (i.e., L0 to L3) and mixed with background noise. A visual inspection of Fig. 4(b) reveals the joint contribution of two individual RTN signals, one switching between $\Delta V_{th}$ levels L0-L1 and the other one switching with lower capture/emission times between L1-L3 and between L0-L2.
2) Identification of $\Delta V_{th}$ trace levels. The next step consists in the application of the WTLP method [5][15] to each recovery trace, in order to identify the number and magnitude of the $\Delta V_{th}$ levels present in the trace. The basic principle of the WTLP technique is the Time Lag-Plot (TLP) method [16], in which the i-th data sample of an experimental signal is represented versus the (i+1)-th data sample as shown in Fig. 5(a). The plot diagonal shows populated data regions corresponding to different data levels (i.e., $I_{DS}$ or $\Delta V_{th}$) while transitions between different levels and background noise are located outside the diagonal.

The TLP methodology can be used when the background noise is low enough, so that the $\Delta V_{th}$ levels are clearly distinguished. However, if the background noise is so relevant that masks the capture/emission transitions, the resulting TLP of the $\Delta V_{th}$ levels are overlapped with the background noise making impossible to accurately locate them. Thus, the improved WTLP [5] is used instead of the TLP to clearly separate the background noise present in the recovery traces from the capture/emission defect transitions. The WTLP represents the probability that a ($\Delta V_{th,i} - \Delta V_{th,(i+1)}$) data sample in a recovery trace corresponds to one of the levels present in the trace. This probability is calculated by the normal bivariate distribution function (1.1) for each data sample. The weighed time lag function defined in equation (1.2) reveals the contribution of all samples in the recovery trace, weighted by its distance to the analyzed $\Delta V_{th,i}$ data sample.

$$\varphi(x,y) = \frac{1}{2\pi\alpha} \exp \left(\frac{-(I_x-x)^2+[I_{th,y} - y]^2}{2\alpha^2}\right)$$

$$\Psi(x,y) = K \sum_{i=0}^{N-1} \Phi_i$$

As shown in Fig. 5(b), where the function $\Psi$ has been plotted against all data samples of the TLP Fig. 5(a), the $\Psi$ function gets higher values in the most populated regions of the TLP. Moreover, those populated regions are located across the main diagonal of the WTLP where each local maximum corresponds to an individual current level present in the recovery trace.

![Fig. 5. Representation of a conventional TLP in an arbitrary example (a) and the corresponding WTLP in (b) [5].](image)

For instance, Fig. 6(a) shows the WTLP resulting from the analysis of trace 1 in Fig. 4(a). In this case, the corresponding $\Delta V_{th}$ variations have been calculated from the comparison between the fresh $I_{DS}$-$V_{th}$ and the $I_{DS}$ current measured after the BTI stress, as explained in step 1. By analyzing the diagonal of the WTLP, four groups of populated data regions, separated in the figure by green dashed arrows, can be distinguished. Each one of the data groups corresponds to a different $\Delta V_{th}$ level present in the recovery trace. Transitions from one data group to the next one are considered as slow defect emissions at a specific $\tau_c$ and the difference between two $\Delta V_{th}$ levels corresponds to the $\eta$ of the discharged defect.

Furthermore, when present, fast capture/emissions on top of each $\Delta V_{th}$ level can also be clearly distinguished, as red-colored regions in the diagonal (i.e., in the $\Delta V_{th}$ levels L0 to L9) with other less populated regions located perpendicularly to the diagonal (i.e., $\Delta V_{th}$ level transitions). Moreover, the ten red populated regions in the WTLP in Fig. 6(a) can be easily associated with each $\Delta V_{th}$ level present in the recovery trace in Fig. 6(b) (pointed out with red arrows).

In order to clearly see the $\Delta V_{th}$ values associated to the RTN transitions detected in Fig. 4(b), a zoom-in of the WTLP (red square in Fig. 6(a)) is plotted in Fig. 7. The zoom clearly shows the 4 red regions across the diagonal, which correspond to the four different $\Delta V_{th}$ levels (i.e., L0, L1, L2 and L3). The blue regions out of the diagonal in the WTLP indicate the

![Fig. 6. (a) WTLP resulting from the analysis of Trace #1 in Fig. 4(a). Corresponding $\Delta V_{th}$ level in the recovery trace for each red-colored zone in the WTLP in (a).](image)
transitions between levels (e.g., L1 → L3). For instance, for t~25s, the signal switches between levels L1 and L3 and later, for t~30s, the signal switches between L0 and L2. These levels result from the joint combination of the two individual RTN signals shown in Fig. 4 (b), coming from the fast charges/discharges of two oxide defects. The results shown in Fig. 7 clearly demonstrate that the application of the WTLP to the $\Delta V_{th}$ recovery trace allows an accurate location of all $\Delta V_{th}$ levels, and is able to distinguish between RTN and BTI contributions.

3) **Background noise removal.** Once the $\Delta V_{th}$ levels are identified with the WTLP method, the procedure assigns the closest $\Delta V_{th}$ level to each sample in the $\Delta V_{th}$ trace. This step quantizes the $\Delta V_{th}$ recovery trace, removing the background noise and keeping only $\Delta V_{th}$ levels associated to captures/emissions of RTN and BTI contributions. For instance, Fig. 8(a)-(c) displays three $\Delta V_{th}$ traces (blue traces), showing high $\Delta V_{th}$ degradation due to the previous stress, and also the $\Delta V_{th}$ trace reconstruction (red traces) with the background noise removed. During the recovery period (in this example, from 2ms to 100s), different $\Delta V_{th}$ levels can be observed because of defect emissions mixed with RTN phenomena, as shown in detail in the zoom-in plots in Fig. 8(b) and (c).

4) **Removal of RTN-related transients.** In order to distinguish between RTN and BTI contributions, we define a square matrix, named Transition Matrix (TM), to store the transitions between different $\Delta V_{th}$ levels in the recovery trace. The dimension of the TM is the total number of $\Delta V_{th}$ levels found, in this particular case ten, each $\Delta V_{th}$ level is denoted as LN where N ranges from 0 to 9. The rows of the TM are defined as the initial $\Delta V_{th}$ level (i.e., i-th $\Delta V_{th}$ level) while the columns are defined as the final ones (i.e., (i+1)-th $\Delta V_{th}$ level). During the $\Delta V_{th}$ trace analysis, three different $\Delta V_{th}$ level transitions can be distinguished:

- **Case (i):** No change of the $\Delta V_{th}$ level, i.e., the $\Delta V_{th}$ level of the i-th data sample equals the (i+1)-th $\Delta V_{th}$ level. For instance, two consecutive $\Delta V_{th}$ samples staying at level L1.
- **Case (ii):** Defect charging. The (i+1)-th $\Delta V_{th}$ level is larger than the i-th $\Delta V_{th}$ level. For instance, two consecutive $\Delta V_{th}$ samples switching from level L1 to L3.

- **Case (iii):** Defect discharging. The (i-th)+1 $\Delta V_{th}$ level is smaller than the i-th $\Delta V_{th}$ level. For instance, two consecutive $\Delta V_{th}$ samples switching from level L3 to L1.

To construct the TM, the $\Delta V_{th}$ recovery trace is swept by its $(i_n, i_{n+1})$ elements and the transition location counter of the TM initial-final $\Delta V_{th}$ level is incremented when needed. Case (i) transitions will lay in the main diagonal of the TM while case (ii) and case (iii) transitions will be located above and below the TM main diagonal respectively. For instance, in Trace 1 in Fig. 8(a), which is actually the same trace in Fig. 4(a), ten $\Delta V_{th}$ levels have been detected by the WTLP method (Fig. 6(a)). The resulting 10x10 TM is filled with all $\Delta V_{th}$ transitions identified during the $\Delta V_{th}$ recovery trace sweep as shown in Fig. 9(a). In order to remove fast RTN transitions from the traces, the method analyzes the data in the TM, distinguishing the following two cases:

- **Slow emission recognition:** This type of defect emissions is characterized by a unique defect discharge to a lower $\Delta V_{th}$ level. In the TM, this appears as a ‘1’ in the corresponding element below the TM main diagonal, and a ‘0’ in the TM symmetric position. Analyzing the TM in Fig. 9(a), three different emissions, without any further capture, can be found marked with blue circles: from initial to final $\Delta V_{th}$ levels L4-L3, L6-L4 and L8-L7.

- **Transients recognition:** The method identifies multiple and consecutive transitions between two distinct $\Delta V_{th}$ levels. For instance, all transitions of the combined RTN signals in Fig. 4(b), uncoupled by the WTLP method in Fig. 7(b), are marked in the TM of Fig. 9(a) with a green dashed square, which correspond to transitions from/to levels L0...
A clear example of a transient in the TM is highlighted with a solid red square box, which corresponds to the RTN signal shown in the inset of Fig. 4(a). The main diagonal (of the red square box) indicates the number of samples at the same $\Delta V_{th}$ level (i.e., initial and final $\Delta V_{th}$ levels L6 or L7) while hundred ten defect discharges (i.e., initial L7 to final L6) and 109 defect charges (i.e., initial L6 to final L7) take place in the RTN signal.

Fig. 10 displays the resulting 1, 2 and 3 traces after the application of the methodology showing a total of 3, 8 and 4 slow emissions, without artifacts coming from fast defect transitions and background noise, respectively.

5) **Defect parameters extraction.** The last step consists in obtaining the $\tau_e$ and $\eta$ parameters of the slow defect emissions. This is done by locating the elements below the TM main diagonal showing single discharge that have a symmetrical zero above the TM main diagonal. The $\eta$ value of each detected defect is obtained by subtracting the final $\Delta V_{th}$ from the initial $\Delta V_{th}$. To allow the evaluation of the $\tau_e$ associated to the defect, when the TM is constructed, also the times at which ($i$, $i+1$) transitions occur are saved. When a slow discharge is encountered, the $\tau_e$ is assigned to the computed $\eta$ value so it the defect is characterized by the tuple $\{\tau_e, \eta\}$.

### III. METHOD VALIDATION

In order to evaluate the validity of the described parameter extraction method, BTI or HCI recovery data can be used. Here, four individual BTI tests, each one involving 248 CMOS transistors with 8 different channel sizes (i.e., W and L), have been considered for defect parameter extraction. Fig. 11 shows a schematic representation of the measurement setup [12] used for the characterization of the ENDURANCE chip devices with the following list of items:

- A full-custom printed circuit board (PCB), where the ENDURANCE IC chip is inserted for DUT measurements.
- The Keysight Semiconductor Parameter Analyzer (SPA) model B1500A.
- An Agilent E3631A power supply for PCB and IC biasing.
- A T-2650BV Thermonics precision temperature system.
- A USB Digital Acquisition System (DAQ), model USB-6501 from National Instruments.
- A personal computer (PC) equipped with Microsoft Windows® operating system and Matlab® software, both using 64-bits architecture.

The Agilent E3631A power supply, the Keysight B1500A SPA and the Thermonics temperature system T-2650BV instruments are connected using an IEEE 488 General Purpose Interface Bus (GPIB), in order to send and receive data during
tests executions. For the communication with the USB-6501 DAQ, a USB 3.0 connection is used.

During the BTI experiments, each device has been tested using a 6-cycle SM scheme, in which the duration of the stress phases is increased logarithmically (i.e., 1s, 10s, 100s, 1000s, 10,000s and 100,000s), while the measurement (i.e., recovery) phase time always lasts 100s. During the stress period, a gate-source voltage \( V_{GS} \) has been applied maintaining the drain-source voltage \( V_{DS} \) at 0V (four different \( V_{GS} \) voltages have been considered), while for measurements \( V_{GS} = V_{th} \) and \( V_{DS} = 100mV \). The total test time required for a 6-cycle SM test on a single device takes 111,111s + (6· 100s) = 111,711s ≈ 31 hours. Thus, for the four BTI tests involving 992 devices, the total test time, with conventional serial testing procedures, would be ≈ 3.5 years, a considerably large, and prohibitive, test time. In this work, and thanks to the array-based IC chip ad-hoc design, in order to significantly reduce the total testing time, the stress phases of each BTI test have been parallelized, without overlapping the measurement phases of any device. In this case, the testing time is reduced to only sixty-four hours per BTI test, resulting in a 4-BTI test that only lasts ten days, a significant test time reduction when compared to conventional serial testing.

From the automated analysis of the data, a total of 10,582 slow emissions have been identified. Fig. 12 shows the histogram plots of the extracted \( \tau_e \) from the defects found during the recovery traces showing that charges are emitted during the very first moments of the recovery phase. Fig. 12 also indicates increasing \( V_{GS} \) voltage during the stress results in an increasing number of slow emissions. Fig. 13 shows the statistical results obtained from the analysis of all \( \{ \tau_e, \eta \} \) tuples extracted by using the method detailed in this work. Fig. 13 (a) shows the dependence of the \( \langle \eta \rangle \) value on the transistor area. The results demonstrate that \( \langle \eta \rangle \) decreases with the transistor area. If the RTN related transients were not removed from the recovery traces before the slow emissions identification, a large number of ‘false’ events (with equal \( \Delta V_{th} \) as the RTN amplitude) would have been taken into account during the \( \langle \eta \rangle \) calculation. Therefore, the resulting \( \langle \eta \rangle \) value, for all tested geometries, would have been close to the \( \Delta V_{th} \) of the fastest RTN, masking the actual \( \langle \eta \rangle \) of the BTI-related defects. Fig. 13 (b) also shows that \( \eta \) is exponentially distributed. The results are in agreement with those in the literature [1], which supports the validity of the methodology.

Fig. 12. \( \tau_e \) histogram plots found during recovery analysis for the 80nm/60nm transistor geometry after 10,000s of pure BTI stress at four different \( V_{GS} \) stress voltages.

**Conclusion**

In this paper, a smart methodology for the massive extraction of CMOS aging parameters has been presented. In order to reduce the testing time required for the massive analysis of the aging of nanometer MOSFETS (from years to days), our fabricated array-based IC chip in combination with a stress parallelization technique have been used. Our method analyzes the BTI/HCl aging recovery traces and looks for \( \Delta V_{th} \) voltage drops related to slow emissions after the application of accelerated stress patterns. The method identifies and removes the fast transient capture/emissions (RTN) and the background noise in the recovery traces to avoid artifacts during the CMOS defect parameters extraction. In the application example, the described method has allowed to extract the \( \{ \tau_e, \eta \} \) tuples of more than 10,000 defects emissions that can be used as input parameters to stochastic aging compact models.

**Acknowledgements**

This work has been supported in part by the TEC2013-45638-C3-R and TEC2016-75151-C3-R Projects (funded by the Spanish Ministerio de Economia, Industria y Competitividad and ERDF). Javier Diaz Fortuny research activity has been supported by the FPI grant BES-2014-067855.
REFERENCES


