New method for the automated massive characterization of Bias Temperature Instability in CMOS transistors

P. Saraza-Canflanca¹, J. Diaz-Fortuny², R. Castro-Lopez¹, E. Roca¹, J. Martin-Martinez³, R. Rodriguez³, M. Nafria³, F. V. Fernandez⁴*

¹Instituto de Microelectrónica de Sevilla, IMSE-CNM (CSIC/Universidad de Sevilla), Sevilla, Spain
²Electronic Engineering Department (REDEC) group, Universitat Autònoma de Barcelona (UAB) Barcelona, Spain
*francisco.fernandez@imse-cnm.csic.es

Abstract—Bias Temperature Instability has become a critical issue for circuit reliability. This phenomenon has been found to have a stochastic and discrete nature in nanometer-scale CMOS technologies. To account for this random nature, massive experimental characterization is necessary so that the extracted model parameters are accurate enough. However, there is a lack of automated analysis tools for the extraction of the BTI parameters from the extensive amount of generated data in those massive characterization tests. In this paper, a novel algorithm that allows the precise and fully automated parameter extraction from experimental BTI recovery current traces is presented. This algorithm is based on the Maximum Likelihood Estimation principles, and is able to extract, in a robust and exact manner, the threshold voltage shifts and emission times associated to oxide trap emissions during BTI recovery, required to properly model the phenomenon.

Keywords—Reliability, Aging, Characterization, Maximum Likelihood Estimation, Time Dependent Variability, Bias Temperature Instability

I. INTRODUCTION

Reliability is a serious concern in analog and digital circuits, especially in deeply-scaled CMOS technologies due to the variability of key transistor parameters such as the threshold voltage (Vth) and the effective mobility (μeff). Time-zero variability (TZV), also known as spatial or process variability, accounts for permanent shifts of device parameters due to imperfections of the fabrication process. Time-dependent variability (TDV) includes transient effects, such as Random Telegraph Noise (RTN), and aging effects, such as Hot Carrier Injection (HCI) and Bias Temperature Instability (BTI), which have become a critical issue for circuit functionality [1]. BTI is gate-voltage and temperature activated and consists in the progressive degradation over time of transistor parameters, e.g., an increase in the absolute value of the threshold voltage, during the circuit operation. The degradation caused by BTI has both a permanent and a recoverable component. The recoverable component recovers when the stress ceases, while the permanent component stays. For technology nodes in the nanometer range, BTI reveals a discrete and stochastic behavior that has been associated to the charge/discharge of defects during stress/relaxation [2], [3].

In order to minimize the impact of BTI on the degradation of circuits, it is necessary to implement reliability-aware design methodologies. To this end, it is fundamental to develop models and simulation tools that correctly describe and predict the phenomenon. However, for transistors with channel dimensions in the nanometer range, models and simulators should account for the above-mentioned stochastic nature of BTI. For this purpose, the Probabilistic Defect Occupancy model has been reported [4]. This model uses a parameter η to describe the change in the threshold voltage Vth of a transistor caused by the trapping of a charge carrier in a defect, and two more parameters, the charge capture and emission times (τc and τe), to describe the occupancy probability of the defect under given bias and temperature conditions. All this, combined with the number of defects per device (N), is used to calculate the Vth shift in transistors under different bias conditions. Using this model, reliability simulation tools for integrated circuits (ICs) have been reported [5]. However, the scenario is not complete just with a model and a simulation tool. To be effective, a characterization of the model parameters is indispensable. Due to the intrinsic stochastic nature of these parameters, a large number of transistors should be characterized to get statistically significant results.

Conventional BTI characterization experiments are based on the application of serialized stress-measurement (SM) sequences to a number of transistors by using a probe station. However, due to the need of analyzing a large number of transistors in order to get statistically sufficient results to characterize BTI in nanometer-scale CMOS transistors, this approach can lead to unaffordable testing times at the laboratory. To overcome this issue, several array-based approaches have been reported [6]-[10]. For instance, the chip in [10] contains more than 3,000 transistors, is equipped with a Force & Sense architecture to avoid voltage drops, and allows to substantially reduce the BTI characterization time by means of a smart stress parallelization technique that allows all the devices to have equal stress-recovery time ratios.

A typical BTI experiment consists in applying successive stress and recovery phases to each transistor. The technological parameters of the device before the stress phase can be characterized by means of a drain current vs. gate voltage (Id-Vg) plot to have some reference values. Then, after the stress has been applied, the recovery current is measured and the discrete current jumps caused by BTI detrapping events are analyzed. These detrapping events lead
to an increase in the absolute value of the current. The impact of BTI aging can be modeled through the study of the emission time ($\tau_e$) and the impact on the total threshold voltage of each defect ($\eta$). Additionally to these detrapping BTI events, RTN trapping/detrapping events may occur. Unlike BTI detrapping events, in which the absolute value of the current increases, the discrete current jumps in RTN have been associated to both charge trapping and detrapping in and from defects, which lead respectively to a decrease and an increase of the absolute value of the drain current. These RTN transitions, together with the noise caused by the on-chip electronic circuitry and the measurement setup itself, can hinder the detection and analysis of the BTI transitions. Taking into consideration this information, an algorithm that aims to correctly analyze a BTI recovery current trace should:

1. Identify the noise-free current levels.
2. Classify the transitions between those levels as RTN- or BTI-induced transitions and remove the former ones.
3. Extract from this noise- and RTN-free trace the parameters of the BTI discharges, namely the emission time and the threshold voltage shift corresponding to each defect.

Apart from the previous considerations, since hundreds or even thousands of BTI recovery traces will have to be analyzed, the designed algorithm should work in an automatic manner, with as little supervision of the user as possible. An algorithm addressing these goals has been reported in [11]. This algorithm works by building a current histogram of the experimental data, fitting it with a function $\psi$, and taking the maxima of this fitted function as the current levels of the BTI recovery trace. Although this method can work well when applied to individual traces, the number and position of the detected current levels strongly depend on the selected number of histogram bins. Therefore, the user must check which is the number of histogram bins that leads to good results for each current trace, which limits the automation of the whole process and significantly increases the data processing time.

In this work, we introduce a novel fully-automated method that fulfills all the previously listed requirements for a correct extraction of the BTI parameters while requiring no intermediate user supervision and intervention.

The rest of the paper is structured as follows. In Section II, the characterization setup used to obtain the BTI current traces is presented. In Section III, the existing method for an automated characterization of the BTI parameters is reviewed and its limitations are highlighted. In Section IV, the novel proposed method is introduced and and experimental results are shown. Finally, conclusions are drawn in Section V.

II. Characterization System

The characterization setup employed to collect the experimental data used in this work consists of a 65-nm CMOS array-based IC chip and a dedicated experimental setup.

1. The characterization chip

The characterization chip reported in [10] was designed to perform both TZV and TDV characterization tests on transistors, including BTI tests. It was fabricated in a 1.2-V, 65-nm CMOS technology and has a chip area of 1.8 x 1.8 mm². The chip contains 3,136 MOS test transistors, half of which are pMOS and the other half nMOS. The number of transistors per chip allows a statistically-significant characterization of BTI that accounts for the stochastic nature of the phenomenon, as well as the study of its voltage and temperature dependences. There are eight different transistor geometries included in the chip to enable the study of the dependence of TDV effects with the channel geometry, being 80 nm x 60 nm the dimensions of the smallest channel. Each device is embedded into a unit cell that includes the necessary circuitry to allow, together with a full custom digital control circuitry (common to the rest of the array), individual access to each device. Each unit cell has a Force & Sense architecture for the drain terminal that enables the accurate application of voltage. Another crucial feature of this chip is the possibility to parallelize the stress phases of the BTI experiments, which allows a significant reduction of the characterization time. To enable this parallelization, the drain and gate terminals of each transistor have separate connections to a stress path and to a measure path. This way, several devices can be simultaneously stressed by connecting their terminals to the stress paths while another device is measured by connecting its terminals to the measurement path. The digital control circuitry mentioned above is used to select which paths each terminal is connected to. To highlight the importance of parallelization, a typical BTI experiment in which 784 transistors are characterized with 4 stress-recovery cycles (stress times of 1s, 10s, 100s, 1,000s and 10,000s, and recovery times of 100s in all cycles) can be considered. This experiment would require 104 days if no parallelization is implemented. With the parallelization scheme of this chip, the same experiment takes only 4 days.

2. Experimental Setup

Fig. 1 displays a schematic representation of the experimental setup used for the BTI characterization of the transistors of the chip. The elements of this setup are:

![Experimental Setup](image_url)
A full-custom printed circuit board (PCB). On-board shielded triaxial connectors allow the access to the analog signal chip pads.

An Agilent E3631A power supply for the biasing of the PCB and the IC.

A Keysight Semiconductor Parameter Analyzer (SPA) model B1500A. This instrument is equipped with 4 High Resolution Sense Measurement Units (HRSMU), each of which provides Force & Sense triaxial outputs for precise voltage application and current measurement.

A T-2650BV Thermonics precision temperature system, which allows the accurate application of temperatures in the range between -40ºC and 170ºC in order to study the dependence of the BTI parameters with temperature.

A USB data acquisition system from National Instruments, which provides the digital signals for the control of the chip.

A IEEE 488.1 GPIB BUS is used for the communication between the controller (a personal computer) and the instrumentation. In order to exploit all the measurement potentialities of the chip, a toolbox for the automation of the characterization tests is used. This software works under the Matlab® programming environment, and is connected to the experimental setup, including the Keysight SPA, the Thermonics temperature system, the Keysight power supply, the National Instruments digital acquisition system and the PCB. This toolbox allows the user to define the desired characterization tests in just a few seconds.

III. THE WTLP-BASED METHOD FOR THE EXTRACTION OF BTI PARAMETERS

To the best of the authors’ knowledge, an algorithm that aims at automatically extracting the parameters needed to model BTI from experimental transistor current traces has only been presented in [11]. This algorithm detects current levels based on the weighted Time Lag Plot (wTLP) method [12], which is an improved version of the Time Lag Plot (TLP) method [13]. The TLP is constructed by plotting the $i$-th point of the current trace in the $x$-axis and the $(i+1)$-th point in the $y$-axis for the full current trace. The points in the diagonal of the TLP ($x = y$) correspond to defined current levels, while the off-diagonal data points correspond to transitions between different current levels. However, the identification of the current levels may be difficult due to the existence of background noise. The wTLP method has been presented as adequate even when the background noise is large [12]. In it, a bivariate normal distribution is defined for each point of the TLP with coordinates $(I_i, I_{i+1})$:

$$\phi = \frac{1}{2\pi \sigma^2} \exp \left\{ -\frac{((I_i-x)^2 + (I_{i+1}-y)^2)}{2\sigma^2} \right\}$$  \hspace{1cm} (1)

where $\sigma$ is the standard deviation of the noise. This distribution represents the probability that the point $(I_i, I_{i+1})$, corresponds to a level or to a transition in the location $(x, y)$ of the TLP space. The weighted time lag function $\psi$ is defined as:

$$\psi(x, y) = K \sum_{i=1}^{N-1} \phi_i$$  \hspace{1cm} (2)

being $K$ a normalization constant chosen so that the maximum value of $\psi$ is equal to 1 and $N$ the number of current data points. The contribution of each point of the TLP to $\psi$ is in this way weighted by the distance between the position of this point and $(x, y)$, so that $\psi$ takes higher values in the most populated regions of the wTLP. This approach generates a very visual result, appropriate for human inspection of the intricacies of the current trace. In Fig. 2, an experimental current trace is displayed together with its corresponding wTLP.

In order to extract more quantitative information, it is possible to build a current histogram and fit it with the diagonal of the obtained $\psi$, denominated wTL $\psi$ diagonal. The diagonal of $\psi$ is built by evaluating (2) for $x = y = x_{\text{hist}}$, where $x_{\text{hist}}$ is the point at which each histogram bin is centered, and summing for all histogram bins. Then, this diagonal is fitted with respect to the histogram by varying its standard deviation $\sigma$. The best match between the wTL $\psi$ diagonal function and the current histogram is obtained when $\sigma$ equals the value of the standard deviation of the background noise by extracting the maxima of the fitted wTL $\psi$ diagonal function, it is in principle possible to obtain the location of the current levels and, thus, the amplitude of the current jumps between these levels caused by BTI or RTN trapping and detrapping events. However, the choice of the bin size of the current histogram can yield different position of the current levels and even a different number of current levels, hence incorrectly determining the number of defects and the amplitudes of the current transitions. For the sake of illustration, using the current trace displayed in Fig.

![Current Trace](image)

**Fig. 2.** Measured recovery current trace in a pMOS transistor of the characterization chip in which only RTN-induced transitions are present (top), and its corresponding wTLP representation (bottom). In the wTLP plot, the current levels can be observed in the diagonal region, while the transitions between these levels can be observed in the off-diagonal regions.
2. when the wTL ψ diagonal function is fitted to a histogram with 60, 100, 140 and 200 bins, 7, 9, 16 and 43 maxima are found, respectively. This means that, depending on the number of bins chosen for the current histogram, 7, 9, 16 and 43 current levels are detected for the same current trace. The histograms with 60 and 200 bins, together with their corresponding wTL ψ diagonal fitted functions, are shown in Fig. 3. Due to this lack of robustness, although this method can lead to good results when the user carefully applies it to individual traces by selecting the appropriate number of histogram bins, it is not adequate to automatically analyze hundreds or thousands of them.

IV. A NOVEL MAXIMUM-LIKELIHOOD-ESTIMATION-BASED METHOD FOR THE AUTOMATIZED EXTRACTION OF BTI PARAMETERS

A flow diagram of the process for the extraction of the BTI parameters, from input files containing the recovery current traces to the output files with the information about the extracted parameters, is displayed in Fig. 4. The parameters that this method aims to extract are the emission time (τₑ) of each BTI defect and its contribution (η) to the total threshold voltage degradation of the transistor (ΔVth).

To this end, the steps followed by the algorithm are:

1. The detection of the M current levels.
2. The elaboration of a clean, noise-free current trace by assigning to each experimental current point the closest of the M levels.
3. The identification of each transition between current levels as a RTN or BTI transition and the consequent elimination of the former ones, this way generating a noise- and RTN-free current trace.
4. The extraction of τₑ and η for each BTI defect from the noise- and RTN-free current trace. There are several methods used to convert the current shift to a threshold voltage shift [14]-[16].

1. Detection of the current levels

Let us consider a device with a number of defects such that the trapping/detrapping mechanisms (since the current trace can have both BTI and RTN events) yield a recovery trace with M current levels. The background noise of the experimental data can be approximated by a Gaussian distribution function that is independent of the different current levels. Therefore, the measured current trace can be reasonably considered as samples of the following probability density function:

\[
f(I|\theta) = \frac{1}{K\sqrt{2\pi\sigma^2}} \sum_{j=1}^{M} A_j e^{\frac{(I-I_j)^2}{2\sigma^2}}
\]

where \( \theta = \{\sigma, A_1, \ldots, A_M, I_1, \ldots, I_M\} \) is the vector of parameters for this density function: \( \sigma \) represents the standard deviation of the background noise, \( I_j \) is the value of each current level and \( A_j \) is its height in the probability density function. Parameter \( K \) is a normalization constant so that the area below the probability density function is unity:

\[
K = \sum_{j=1}^{M} A_j
\]

Let us consider \( N \) samples of the current trace \( \{I_1, \ldots, I_N\} \). It can be naturally assumed that all samples are independent and identically distributed; then, the joint density function of all observations is:

\[
f(I_1, \ldots, I_N|\theta) = \prod_{i=1}^{N} f(I_i|\theta)
\]

The likelihood of a set of parameter values \( \theta \), given the observed results, is equal to the probability of obtaining those results as a function of \( \theta \), i.e.,

\[
L(\theta) = \prod_{i=1}^{N} f(I_i|\theta)
\]

The Maximum Likelihood Estimation (MLE) method is used to estimate the parameters of a statistical model given the known outcome: \( \{I_1, \ldots, I_N\} \). The Maximum Likelihood Estimation aims at identifying the parameter values \( \theta \) that make the observed data \( \{I_1, \ldots, I_N\} \) the most probable, i.e., that maximize (6). Since the number of samples \( N \) can amount to many thousands, it is more convenient to work with the natural logarithm of (6), and, in this way, the multiplication in (6) can be converted into a summation. This implies no difference in terms of finding the maximum of (6) since the logarithm is a monotonically increasing function.

Fig. 3. Current histograms with 60 and 200 bins corresponding to the current trace displayed in Fig. 2, along with their fitted wTLP diagonal ψ. The wTLP diagonal method leads to the detection of 7 and 43 current levels when current histograms with 60 and 200 bins are used, respectively.
Different optimization algorithms can be applied to this maximization problem. For efficiency reasons, a deterministic local search method is applied in our case. The convergence of this method depends on an appropriate starting point. The location and amplitudes of the peaks of the histogram are good starting points to this end. In order to avoid the identification of any other type of background noise as RTN or BTI transitions, a minimum distance between histogram peaks is established. This way, if there are two peaks in the current histogram within a distance considered to be smaller than the experimental noise, only the largest of them is considered for the starting point of the optimization algorithm. Notice that, unlike the wTLP method, the density function is obtained from the optimization of (6), and not by fitting the function to the current histogram.

2. Generation of a clean or noise-free current trace

Once the M current levels have been identified, a clean or noise-free trace can be generated by just assigning to each experimental current point the value of the closest of the M levels. Two examples of experimental BTI recovery current traces with their corresponding calculated noise-free traces are displayed in Fig. 5 and 6. These recovery traces were measured at $V_{GS} = 0.6$ V and $V_{DS} = 0.1$ V after the respective transistors had been stressed at $V_{GS} = 2.5$ V and $V_{DS} = 0$ V. To illustrate the strength of the algorithm, Fig. 7 displays a complex multi-level experimental current trace. The clean traces for Fig. 5, 6 and 7 have been generated without any changes by the user to the optimization parameters. This robustness is fundamental when hundreds or thousands of recovery traces must be analyzed, so that the whole process can be automated, thus enormously reducing the amount of time required.

3. Elimination of RTN transitions

Once a noise-free current trace has been obtained, the next step is to identify each current jump as an RTN or a BTI event in order to eliminate the former ones. Transition events in which the absolute value of the current decreases can be straightforwardly assigned to RTN, since detrapping BTI events always cause an increase in the absolute value of the current. Events in which the absolute value of the current increases are assigned to RTN detrapping if an inverse RTN event (trapping) of the same amplitude occurs. In Fig. 5, there are only 5 BTI detrapping events present. In Fig. 6, there are both RTN and BTI events, specifically 3 BTI...
detrapping events and abundant RTN trapping and detrapping transitions caused by a single RTN defect. These RTN events can then be eliminated to obtain a noise- and RTN-free current trace. Such a trace for the example displayed in Fig. 6 is shown in Fig. 8.

4. Extraction of $\tau_e$ and $\eta$ for each BTI defect

Once the RTN traces have been cleaned, an output file is generated with the information about each of the BTI discharges detected, namely its current amplitude and its emission time. The emission times can be directly plugged into the Probabilistic Defect Occupancy model. The amplitude of the current jump is converted into a threshold voltage shift $\eta$ by applying the method in [16]. For the sake of illustration, the parameters extracted by the algorithm in an automated manner for the experimental traces in Fig. 5 and Fig. 6 are presented here. The recovery current trace displayed in Fig. 5 has 5 BTI detrapping events, whose parameters are $\tau_{e1} = 0.01$ s, $\eta_1 = 8.5$ mV; $\tau_{e2} = 18.33$ s, $\eta_2 = 5$ mV; $\tau_{e3} = 20.31$ s, $\eta_3 = 2$ mV; $\tau_{e4} = 61.09$ s, $\eta_4 = 7$ mV, and $\tau_{e5} = 92.19$ s, $\eta_5 = 2$ mV. The current trace displayed in Fig. 6 has 3 BTI detrapping events, whose parameters are $\tau_{e1} = 19.91$ s, $\eta_1 = 1.5$ mV; $\tau_{e2} = 35.48$ s, $\eta_2 = 5$ mV, and $\tau_{e3} = 49.72$ s, $\eta_3 = 2$ mV.

V. CONCLUSION

In this work, a novel algorithm based on the Maximum Likelihood Estimation method to automatically analyze the BTI current traces and extract their parameters has been presented. The robustness of this algorithm allows the user to apply it to hundreds or thousands of different current traces without the need of individually inspecting each of them in order to select the best conditions for the optimization, which is a fundamental feature for the automation of the process.

REFERENCES