Flexible Setup for the Measurement of CMOS Time-Dependent Variability With Array-Based Integrated Circuits

Javier Diaz-Fortuny, Pablo Saraza-Canflanca, Rafael Castro-Lopez, Elisenda Roca, Javier Martin-Martinez, Rosana Rodriguez, Francisco V. Fernandez, and Montserrat Nafria

Abstract—This paper presents an innovative and automated measurement setup for the characterization of variability effects in CMOS transistors using array-based integrated circuits (ICs), through which a better understanding of CMOS reliability could be attained. This setup addresses the issues that come with the need for a trustworthy statistical characterization of these effects: testing a very large number of devices accurately but, also, in a timely manner. The setup consists of software and hardware components that provide a user-friendly interface to perform the statistical characterization of CMOS transistors. Five different electrical tests, comprehending time-zero and time-dependent variability effects, can be carried out. Test preparation is, with the described setup, reduced to a few seconds. Moreover, smart parallelization techniques allow reducing the typically time-consuming aging characterization from months to days or even hours. The scope of this paper thus encompasses the methodology and practice of measurement of CMOS time-dependent variability, as well as the development of appropriate measurement systems and components used in efficiently generating and acquiring the necessary electrical signals.

Index Terms—Aging, automated characterization lab, bias temperature instability (BTI), device modeling, hot carrier injection (HCI), random telegraph noise (RTN), variability.

I. INTRODUCTION

I N THE last decades, CMOS technologies have experienced a steady scaling predicted by Moore’s law, hugely improving the device performances. In this dive into the nanometer scale, dealing with (time-zero) variability (TZV) has been a constant struggle [1]. However, while manufacturing process variations have been kept under control (by steadily innovating in manufacturing tools and random variations (such as random discrete dopants, line edge roughness, metal grain granularity,) arising from the discreteness of charge and matter granularity have become a critical concern for circuit design in advanced nodes since they cause fluctuations that result in transistor electrical parameter mismatches between identically designed devices (e.g., threshold voltage mismatch) [2], [3]. Though (planar or 3-D) SOI technologies mitigate this problem, it remains an important concern [4], [5].

Another source of transistor variability, known as time-dependent variability (TDV), appears during circuit operation and can be temporary or permanent. TDV phenomena, like bias temperature instability (BTI) [6], hot carrier injection (HCI) [7], [8], or time-dependent dielectric breakdown (TDDB) [9], which have been intensively investigated in the last decades, are reflected in the progressive degradation of the transistor parameters over time. All these phenomena become intrinsically stochastic at the nanometer scale. Another TDV phenomenon is the random telegraph noise (RTN) [10], [11] transient effect, which consists in current fluctuations between two or more differentiated levels.

Today, all TDV phenomena are a serious concern because of their inherent variability impact on circuit and system variability in advanced integration nodes [12], [13]. Therefore, statistical modeling and characterization of the device degradation produced by these sources of variability is mandatory. This degradation has been captured in several models reported so far, such as the probabilistic defect occupancy (PDO) model [14], representing a key step in variability-aware design methodologies [15]. One major challenge related to the intrinsic random nature of the variability mechanisms is that a large number of transistors should be characterized to get statistically significant results. Thus, when thousands of devices should be tested, traditional on-wafer device testing techniques based on serial device characterization result in unfeasible testing times. In this scenario, integrated circuits (IC) with transistor arrays have been proposed for massive device characterization of TZV [16] and aging (BTI/HCI) [17]–[23] to reduce test times significantly. Array-based ICs for RTN characterization have also been developed [24], [25]. Recently, the first array-based IC enabling the characterization of all these effects has been proposed [26], [27].

Despite the advantages over conventional wafer-based techniques, array-based solutions bring new challenges (e.g., IR drops on internal chip metal lines, large number of devices,
or individual device access) critical not only for the design of the IC itself but also for the way in which it will be tested. Therefore, efficient and accurate setups must be developed to provide users with a hardware and software solution for fully-automated device-level variability characterization.

In this scenario, this paper presents a flexible characterization setup that consists in a complete laboratory instrumentation system and a fully automated characterization software, together with a graphical user interface (GUI), to fulfill the hardware and software requirements for a trustworthy and straightforward transistor characterization.

The rest of this paper is organized as follows. A review of variability phenomena affecting CMOS transistors, along with the suitable characterization procedure, is explained in Section II. Section III briefly shows the features of the fabricated 65-nm CMOS array-based chip for which the measurement setup has been developed. A detailed description of the measurement hardware elements employed for the massive characterization of the chip devices is given in Section IV. Section V describes the design of the characterization software. Illustrative results obtained from the chip with the proposed setup are presented in Section VI and, finally, the conclusion is drawn in Section VII.

II. VARIABILITY CHARACTERIZATION IN CMOS TRANSISTORS

A. I–V characterization

For the characterization of transistor variability, it is critical to obtain the transistor parameters, such as threshold voltage ($V_{th0}$) or mobility ($\mu_0$), in order to assess their variations. The transistor parameters are extracted from the drain current versus gate–source voltage ($I_{DS}$–$V_{GS}$) and the drain current versus drain–source voltage ($I_{DS}$–$V_{DS}$) characteristics.

B. Random Telegraph Noise characterization

Performance of short channel devices is strongly influenced by a random switching of $I_{DS}$ between two or more levels along time. These current fluctuations are related to the capture and emission of charge carriers by oxide and interface traps, showing a large dependence on device biasing and temperature. The RTN phenomenon has been recognized as a significant variability source since it is responsible for device parameter shifts that increase inversely with area scaling [28].

The standard RTN characterization procedure consists in continuously measuring the $I_{DS}$ current while maintaining a constant $V_{DS}$ and $V_{GS}$ bias voltage. Sufficient accuracy of the current measurement is required since the RTN-induced current variation may amount to just a few nano amps. The time required for a statistically sound characterization of RTN depends on the trap(s) intrinsic capture/emission time constants, which span several orders of magnitude. Most previous efforts have been directed to RTN analysis methods (see [29]) or to the necessary timings for a meaningful estimation of time constants [30]. Although these aspects deserve attention, only recently array-based solutions devised to satisfy the required statistical characterization have been reported [24]–[27]; bringing new challenges, e.g., automation, individual device access, or accurate control of bias voltages for an accurate characterization of the bias voltage dependence of time constants.

C. BTI and HCI Characterization

BTI and HCI are both aging phenomena that produce a gradual shift of the device parameters over time: mainly an increase of the threshold voltage ($V_{th0}$), but also a decrease in the channel mobility ($\mu_0$). Similar to RTN, the capture and emission of charges to/from traps are believed to be behind these shifts, with charges being stochastically trapped when a certain voltage is applied (therefore increasing the threshold voltage), and emitted when the voltage decreases (therefore the threshold voltage starts recovering its original value). Such shifts depend on the local temperature, as well as on the (time-varying) voltages applied. The shifts can lead to severe, even fatal, circuit performance degradation. Therefore, in order to predict and mitigate these effects, the study of BTI and HCI phenomena has a renewed interest in nanometric technologies [31], [32]. Proper modeling [14], model parameter extraction [33], [34], and simulation strategies [35]–[37] are essential not only to implement appropriate variability-aware design techniques but also for proper design of runtime compensation strategies [15], [38].

BTI/HCI transistor characterization commonly uses a stress-measurement (SM) technique that allows extracting the degraded transistor parameters and computing their shifts from their initial pristine values [6]. The SM technique used in this paper follows three steps:

1) Initial Device Characterization: Measurement of the pristine $I_{DS}$–$V_{GS}$ curve (see Section II-A) to extract the device parameters before the application of any kind of stress.

2) Stress: BTI and HCI are long-term aging processes that take place while devices operate within their nominal operating voltages. Thus, in order to be able to characterize both aging processes in relatively short and affordable times, the devices must be subject to accelerated stress processes that consist in the application of high temperatures and/or overvoltages, e.g., twice the nominal supply voltage, to the device terminals [7], [39]–[41]. For BTI characterization, a high $V_{GS}$ and $V_{DS} = 0$ V are set, whereas, for HCI, $V_{DS}$ is set at high values.

3) Measurement: The evolution of the drain current with time $I_{DS}(t)$ is registered right after the stress is removed, the exact moment at which the parameters of the stressed devices start recovering. During the $I_{DS}(t)$ measurement, the discharges (i.e., recoveries) of the traps previously charged in the stress phase are measured and characterized. The sampling rate during the measurement must be as high as possible in order to capture fast current events that may occur after the stress phase. At the end of each $I_{DS}(t)$ measurement, an additional $I_{DS}$–$V_{GS}$ characterization is carried out for a complete analysis of the transistor parameter degradation.
As the device parameter degradation depends on the stress time, several SM cycles are applied sequentially to the samples, where the time on each stress cycle is typically increased exponentially. On the other hand, the elapsed time between the stress (second step) and measurement (third step) should be as short as possible and preferably equal for all tested devices because, as soon as the stress is removed, the recovery behavior of the threshold voltage begins.

The characterization of the different phenomena presented in this section, together with the adopted array-based strategy, leads to a set of instrumentation and measurement challenges that the experimental setup should address:

1) Massive characterization of many devices is needed due to the stochastic nature of TDV in nanometer-scale technologies.

2) Transistors with different channel geometries should be characterized to study the dependence of TDV phenomena with device size.

3) Accurate electrical characterization (which results in the need for canceling leakage currents, applying accurate voltages to the device terminals, minimizing the noise in the measurement lines, and ensuring stable biasing) is required in order to correctly study the dependence of TDV with the bias conditions.

4) Accurate thermal characterization over a broad range of temperatures is essential since the TDV phenomena strongly depend on temperature.

5) Characterization times must be affordable.

6) Due to the large number of devices and the long characterization times, the control of the test process and data analysis must be fully automated.

These requirements are addressed by the combined power of an experimental setup composed of a hardware part (Section IV) and a software part (Section V) and the array-based IC, named ENDURANCE, reported in [26] and [27], and whose major characteristics impacting the experimental setup are reviewed in Section III.

III. TEST STRUCTURES

The ENDURANCE chip is capable of characterizing TZV, BTI, HCI, and RTN phenomena [26], [27]. The chip was fabricated in a 1.2 V, 65-nm CMOS technology with an area of $1800 \times 1800 \ \mu m^2$. It includes nMOS and pMOS transistors arranged in two independent arrays, with a total of 3316 regular threshold voltage transistors, hereafter known as the devices under test (DUT). Transistors with different width and length values have been included.

Each DUT of the arrays is surrounded by a digital control circuitry, conforming what is called a unit cell of the array. The control circuitry determines how the terminals of the DUT (gate and drain terminals) are connected to the external chip pads.

The chip, which is encapsulated in a JLCC68 package, includes 64 pads, 36 corresponding to analog signals that connect to the DUT terminals and 28 pads corresponding to digital signals that are connected to the control circuitry of the unit cells. Each array is subdivided into two subarrays that have independent analog signals connected to external pads. Three different paths connect the DUTs with the analog pads: a measurement path, a stress path, and a standby (SB) path, and the gate and drain terminals of each DUT are connected through switches or transmission gates (TG) to all three paths. The drain terminal has force and sense [42] connections for both the measurement and the stress paths, so a total of five analog signals are connected to this terminal. Since negligible current flows through the gate terminal, it has no force and sense connection, so only three analog signals are connected to this DUT terminal.

The switches connecting the three paths are controlled by the digital control circuitry of the unit cell. Depending on the value of the digital control signals applied, each unit cell will be set in one out of the following four operation modes (OM):

1) **Measurement Mode**: It is designed to measure $I$–$V$ characteristics, RTN transient effects, and post-stress BTI/HCI phenomena. For these tests, only low voltages (i.e., from 0 to 1.2 V) are allowed. The physical paths that connect to the DUT terminals are the drain measure path (designed with force and sense lines to compensate for the IR drops) and the gate measure (GM) path.

2) **Stress Mode**: It is designed to apply an overvoltage, i.e., from 1.2 to 3.3 V, to the gate and/or drain terminals during the stress phase of BTI and HCI tests. The physical on-chip paths for this mode are the drain stress path (designed with force and sense lines) and the gate stress (GS) path. The availability of separate stress and measurement analog paths for the drain terminal enables the implementation of smart parallelization schemes, as detailed in Section V.

3) **StandBy Mode**: This mode sets all voltage differences of the DUT terminals at 0 V when neither measurement nor stress is conducted, and is intended to avoid any device degradation while other devices are being measured or stressed.

4) **Off Mode**: This mode opens all switches or TGs of the unit cell, disconnecting the DUT from the analog paths. The off OM serves as a bridge between the SB mode and the measure/stress modes to prevent short-circuits. It is also used to disconnect nonfunctional DUTs.

To avoid aging or degradation of the eight TGs included in each unit cell due to the overvoltages applied during the stress mode, these are designed with IO transistors capable of withstanding voltages up to 3.3 V without significant degradation during the experiments.

IV. HARDWARE SETUP

An experimental setup has been developed for the characterization of the ENDURANCE chip. A schematic representation of this measurement setup is shown in Fig. 1. It consists of the following elements:

1) A full-custom printed circuit board (PCB), where the ENDURANCE IC is inserted for DUT measurements.

2) The Keysight Semiconductor Parameter Analyzer (SPA) model B1500A. The system has four high-resolution sense measurement units (HRSMU), with force and
sense outputs for precise voltage application and current measurement [42]. The analog signal pads are connected to the HRSMUs so that the voltage can be set, and, if necessary, the currents can be measured [43].

3) The Agilent E3631A power supply for PCB (and subsequently IC) biasing using a 5 V/1A dc biasing voltage [44].

4) The T-2650BV Thermonics precision temperature system with a temperature ranging from room temperature to 120 °C [45].

5) A USB Digital Acquisition System (DAQ), model USB-6501 from National Instruments, equipped with 24 digital IO channels provides the digital signals for the IC control [46].

6) A personal computer (PC) equipped with MATLAB.

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

The PCB has been designed to provide easy access to the chip but preserving the accuracy of the measurements. Fig. 2 shows the PCB layout composed of three main blocks: the input digital block (Fig. 2, Section IV-A), the biasing circuitry block (Fig. 2, Section IV-B), and the socket and connectors block (Fig. 2, Section IV-C). These blocks are explained in more detail below.

A. Biasing Circuitry Block

Three precise and stable biasing voltages are needed for the ENDURANCE chip. For this purpose, a dedicated biasing circuitry that uses three CMOS high precision voltage regulators has been implemented in the PCB circuitry (Section IV-B in Fig. 2). This circuitry provides, from a single 5 V–1 A dc power supply voltage, the biasing voltages of all digital circuits of the chip (1.2 V), of the digital IO pads (2.5 V), and of the TG and analog pads that enable stress voltages up to 3.3 V.

B. Input Digital Block

This block (located in Section IV-A in Fig. 2) is in charge of receiving the set of digital signals generated by the DAQ for the chip control. These signals are connected to the PCB through a 32-pin insulation-displacement connector (IDC). A set of five digital electromagnetic isolator chips is used to transfer these digital signals to an isolated PCB power domain where the IC is located so that any noise coming from the PC power domain is reduced. This improves the set-up capabilities, especially for low-level current measurements.

C. Socket and Connectors Block

This block of the PCB includes a 68-pin zero-insertion force socket, where the ENDURANCE chips are inserted for testing, and different triax connectors that provide access to the output signal pads of the four arrays through the B1500 HRSMUs triaxial connectors and cables. The socket allows a large number of chip insertions, ensuring the durability of the test system.

As mentioned in Section III, each DUT terminal (i.e., drain and gate terminals) is connected to three different paths that can be accessed from independent analog pads for each array. SB pads are physically connected to the analog ground or analog power supply voltage, depending on the type of DUT of the subarrays (n- or p-type transistors, respectively). The other pads are accessible through triaxial connectors. For those pads where current has to be measured, i.e., the drain measure force output pads of the four subarrays (DMF_N1, DMF_N2, DMF_P1, and DMF_P2 in Fig. 2), a built-in guard has been implemented between their triaxial connector and the IC socket pin. This guard is connected to the guard shield provided by the triax cables of the B1500 SPA, extending the isolation of the measured current to the PCB. This built-in guard has not been implemented in the other output pads because the current through them is negligible (which happens with the GM, drain measure sense (DMS), and drain stress sense (DSS) lines), or are used for stressing the DUTs and the current is not measured (which happens with the GS and drain stress force (DSF) lines). These signals are physically multiplexed along the four DUT subarrays so that only five additional triaxial connectors are necessary, as seen at the right side of the PCB in Fig. 2.
Finally, a squared area surrounding the IC socket has been left free of connectors. This area is used to place the head of the Thermonics system over the chip so that temperature inside the head (and therefore, the temperature of the chip) can be controlled.

V. SOFTWARE TOOL

For the tests described in Section II, thousands of properly scheduled instructions are needed to control the hardware setup described in the previous section. Therefore, the need for automating the generation of instructions for the control of the hardware setup becomes patently clear. To this end, a software toolbox that generates all necessary GPIB and DAQ instructions named TARS (A Toolbox for Statistical Reliability Modeling of CMOS Devices) has been implemented using MATLAB scripting language [47]. This toolbox features several user-friendly GUIs (illustrated in Fig. 3) to facilitate the definition of measurement tests.

A. Architecture

The transformation of the test parameters specified by the user in Fig. 3 into a set of GPIB commands for instrumentation control, as well as DAQ commands for IC control, follows a modular approach with three main modules and the scripts for the communication between them. These modules are:

1) **Test Plan Generation**: This module receives information about the user-defined tests and generates a script containing a specific test plan. This plan consists of a proper sequence of the electrical tests to be applied to each DUT and the range of test conditions.

2) **Elementary Instruction Generation**: This module transforms each command of the test plan into the set of elementary actions that implement such command.

3) **Test Execution**: This module reads each elementary instruction and generates and sends the corresponding GPIB commands to the instrumentation and the IC control instructions to the DAQ.

These three modules are described in detail in Sections V-C–V-E. But before doing so, it is important to bring out a major advantage of array-based ICs over conventional probe stations: the possibility of executing parallel stress techniques to characterize HCI and BTI of hundreds of devices. However, it is not sufficient that an array-based IC enables that possibility, but more importantly, that the experimental setup must accurately and efficiently exploit such capability. The setup described in this paper uses the stress parallelization technique named all-DUT-parallel-stress-pipeline measurements (ADPSPM), which drastically reduces the total aging test time by stressing devices in parallel while maintaining sequential DUT measurement. This technique is directly involved in the generation of the appropriate test plan and, hence, it is described in Section V-B.

B. Stress Parallelization

Parallelization techniques are used to dramatically speed up the aging tests. In this regard, the stress can be parallelized but only one DUT can be measured at a time to assess its degradation. The PSPM technique [48], [49] deals with this constraint by delaying the SM sequence of each DUT with respect to the previous one, resulting in a “place-and-check” algorithm that partially executes simultaneous (parallel) stress phases and pipeline measurement phases depending on the duration of the first stress period. This is done by delaying a complete number of SM cycles, which essentially means repeating the PSPM tests in series until all DUTs are tested. Therefore, the PSPM technique becomes inadequate for certain values of the initial stress time and number of SM cycles.

The ADPSPM parallel testing technique utilizes SB periods that are introduced between certain measurement and stress phases to make the necessary room to accommodate any number of DUTs. Obviously, the length of the SB periods must be accounted for during the post processing of measurement data to generate model parameters since the devices may continue their recovery during such period. As depicted in Fig. 4, in order to achieve an accurate and efficient SM parallel test, the algorithm handles each test cycle individually. In the first cycle (C1 polygon in Fig. 4), the SM pattern of each DUT is delayed, like in the PSPM approach, in order to ensure that the measurement phases in all DUTs are pipelined. Once all the SM patterns of cycle C1 have been temporarily
allocated, the algorithm starts the distribution of the next cycle (C2 polygon in Fig. 4). The ADPSPM algorithm ensures that the last measurement phase of cycle C1 (for DUT#5 in Fig. 4) does not overlap with the first measurement phase of cycle C2 (for DUT#1), shown in Fig. 4 with a dashed arrow, by inserting the necessary SB periods. Thus, the ADPSPM technique can accommodate any number of DUTs.

To illustrate the advantages of this technique, it is interesting to perform a study of the time complexity of the characterization process. Fig. 5 shows the characterization time per device as a function of the number of DUTs for three different combinations of the number of SM cycles, measurement time, and duration of the stress periods. The measurement time of the SM cycles is 100 s in all cases. The three cases use five, four, and three SM cycles, being the duration of the first stress period 1, 10, and 100 s, respectively, and exponentially increasing in the successive cycles. The blue diamonds represent the time complexity of the SERIAL test, a brute-force approach in which the devices are characterized one at a time by applying the full sequence of SM cycles. The test of the following DUT does not start until all SM cycles of the previous device have finished. This implies that the test time per device amounts to the sum of stress times and measurement times of all SM cycles. If we consider the place-and-check PSPM technique for the conditions shown in Fig. 5(a) in red squares, no parallel stress can be applied because the initial stress time of 1 s only allows to test one single device at a time, thus overlapping with the SERIAL testing time [see inset in Fig. 5(a)]. This means that no time improvement when compared with SERIAL test can be achieved. On the contrary, the ADPSPM technique clearly shows a fast decrease of the test time per device when increasing the number of devices, as shown with the orange circles. In Fig. 5(b), the initial stress time is set to 10 s, allowing the PSPM technique to parallelize only two devices, whereas in Fig. 5(c), the initial stress time is set to 100 s and up to ten devices can be stressed in parallel with the PSPM technique. Correspondingly, the test time per device is reduced to one-half or one-tenth of the SERIAL procedure, but this two-device or ten-device parallel scheme must be repeated until all devices are measured. Also in these cases, the ADPSPM technique reveals a significant test time reduction compared to the SERIAL and PSPM techniques by parallelizing the stress phases whenever possible while maintaining the measurement channel occupied the maximum time (see Fig. 4).

As stated above, only one DUT can be measured at a time. Therefore, even if we ignore the stress times and the necessary parallelization, just the measurement periods imply the product of number of cycles and the measurement time per cycle. This IDEAL time is indicated as a continuous black line in Fig. 5. It represents an unreachable limit of the test time per device. It can be checked in Fig. 5 that in all test situations, the ADPSPM technique tends asymptotically to the IDEAL time when the number of DUTs increases.

C. Test Plan Generation

From the user-specified tests, e.g., those shown in Fig. 3, a test plan must be built. The test plan contains information about each test performed to each DUT. This information includes the DUT identification, the corresponding test, i.e., $I_{DS}-V_{GS}$, $I_{DS}-V_{DS}$, RTN, or BTI/HCI, and the variables for the particular tests, i.e., delay time, voltages, temperature, etc. The test plan is written in a script, whose possible codes are displayed in Table I for each type of test.

For the generation of the test plan, the algorithm described by the flow diagram of Fig. 6 is executed. The algorithm is divided into three different branches: the $I-V$ measurement branch, which is dedicated to $I_{DS}-V_{GS}$ and $I_{DS}-V_{DS}$ test plan generation, the RTN test branch, and the BTI/HCI aging branch, which applies the ADPSPM technique to optimally allocate stress, measurement, and standby periods for
TABLE I

| Test plan columns | Electrical test | #1 (s) | #2 (integer) | #3 (ASCII) | #4 OM (ASCII) | #5 OM code (integer) | #6 (V) | #7 (V) | #8 (V) | #9 (V) | #10 (integer) |
|-------------------|----------------|--------|--------------|------------|---------------|---------------------|--------|--------|--------|--------|--------------|
| I_DS–VGS          | Time           | DUT #  | PMOS/NMOS    | ME         | 33            | V_0                  | Final V_0| --     | Step V_0|
|                  |                |        |              |            |               | V_0                  | V_0    | --     | --     | --     |              |
| RTN              | Time           | DUT #  | PMOS/NMOS    | ME         | 20            | Initial V_0          | Final V_0| Initial V_0| Final V_0| Final V_0|
|                  |                |        |              |            |               | V_0                  | V_0    | --     | --     | --     |              |
| HCI               | Time           | DUT #  | PMOS/NMOS    | ME/ST      | 22/11         | V_0                  | V_0    | --     | --     | --     |              |
|                  |                |        |              |            |               | V_0                  | V_0    | --     | --     | --     |              |
| Stand-by         | Time           | DUT #  | PMOS/NMOS    | SB         | 0             | V_0                  | V_0    | --     | --     | --     |              |

Elementary instruction columns

| GPIB/DAQ variables | Test time | DUT # | PMOS/NMOS | OM | OM code | V_0 | V_0 |
|--------------------|-----------|-------|-----------|----|---------|-----|-----|

For the sake of illustration, let us consider a test example with just five devices where the test plan consists of: 1) measuring the $I_{DS}$–$V_{GS}$ curve of all devices; 2) executing an RTN test with a single $V_{GS}/V_{DS}$ condition and a measurement time of 100 s, for all selected devices; and finally 3) running a five-cycle SM BTI measurement with measurement time of 100 s and exponentially (base 10) increasing test times starting at 1 s using the ADPSPM parallelization algorithm. The resulting test plan script, shown partially in Fig. 7, contains the information that defines all these tests. Lines 2–5 show an excerpt of the $I_{DS}$–$V_{GS}$ test plan. Line 2 sets DUT#1 in measurement mode with a sweep in $V_{GS}$. Then, the DUT is set in SB mode so that it does not suffer any degradation. Notice that the delay time in line 3 (first column) is 0, meaning that the following script line can be immediately executed. The same sequence is applied to the second DUT (lines 4 and 5) until we complete all devices. Lines 9–12 are an excerpt of the RTN characterization. Each device is sequentially measured for 100 s and set at SB mode after each measurement. For the implementation of the test plan of the BTI characterization experiment, the ADPSPM algorithm in Section V-B is executed. Lines 16–38 are an excerpt of the first BTI script lines of the test plan. Line 16 sets the first device in stress mode for 1 second. Then, the device must undergo a 100 s measurement period. But 1 s before the measurement stage finishes, DUT#2 is set in stress (line 18). After the measurement of DUT#1 finishes, it is set in SB (line 20). Thanks to setting DUT#2 in stress...
mode, just the required stress time before the measurement of DUT#1 finishes, DUT#2 is ready for measurement just after DUT#1. The same procedure is sequentially applied to the rest of the devices. When the allocation of the first SM cycle finishes, then, the second one starts with the first device. Notice that the stress period of the second cycle is 10 s. Therefore, 10 s before the measurement of DUT#5 finishes (line 33), the stress of DUT#1 is activated (line 34). This way, when the measurement of DUT#5 concludes, also does the stress time of DUT#1 and this device is ready for measurement (line 38). Therefore, we achieve two goals: first, maximum efficiency as the setup is never idle because it is measuring for the longest possible time, and, second, short and equal delays between stress and measurement.

D. Elementary Test Instruction Generation

The purpose of the test plan is to establish the set and sequence of test actions that have to be performed. The test actions have to be executed through the GPIB commands controlling the different instruments and the control commands of the DAQ for properly selecting and setting the DUTs. However, each line of the test plan script may involve different setting conditions that have to be sequentially established, e.g., each sampling point of the $I_{DS}-V_{GS}$ curve of the test plan line is defined by a different gate voltage. Therefore, this second module transforms the test plan script into a set of sequential elementary instructions. Fig. 8 shows an excerpt of the elementary instruction script obtained from the test plan in Fig. 7.

E. Test Execution

The test execution module, illustrated in the flow diagram in Fig. 9, reads and executes the elementary test instruction script line by line, converting the test variables into the corresponding GPIB and DAQ commands, and sending these to the measurement instruments and the data acquisition system. The engine first analyses if the currently selected DUT must be changed. If this is the case, the module sends the corresponding row and column matrix addresses of the new DUT to the serial chip interfaces through the DAQ. Then, it activates the selected OM by connecting the drain and gate DUT terminals to the required analog IC paths. The same operation is performed if the DUT is not changed but a new OM code must be applied to the current DUT.

After the appropriate DUT is activated and the OM is set, the test execution module extracts the $V_{GS}$ and $V_{DS}$ voltages and sends the appropriate commands to start a stress (ST) or a measurement (ME) phase:

1) Stress: The stress biasing voltages are set to the drain and gate HRSMUs and no current is measured. The stress biasing voltages selected for the first time will be kept until the end of the selected test, ensuring that all DUTs connected in parallel are stressed in the same
conditions and without interruptions during the complete test duration.

2) Measurement: The biasing voltages of this mode can be changed each time a DUT is set into the measurement OM because only one DUT will be measured at any time. The test execution protocol executes the measurement in four steps to have a minimum time gap between ST and ME phases and to compensate the current leakage of the IC for the particular test conditions, i.e., DUT biasing and physical location. These four steps are as follows.

a) PAD Biasing: The algorithm biases the drain and gate measure lines (DMF, DMS, and GM in Fig. 2).

b) Current Calibration Through Leakage Measurement: A leakage current measurement is performed through the DMF path to capture the current leakages happening in the path to the DUT drain (especially those coming from the TGs in that path). This leakage current will be used to calibrate the DUT current measurement.

c) $I_{DS}$ Channel Measurement: The current through the measurement path is captured while the DUT still remains in its previous OM (i.e., stress OM), before digitally changing the OM code of the DUT unit cell. At this step, the drain/gate DUT terminals are still connected to the stress path. This step is obviously skipped if there was not a previous ST phase, e.g., RTN or $I–V$ characterization.

d) $I_{DS}(t)$ DUT Measurement: After starting the measurement, the OM code of the DUT unit cell is digitally changed to ME mode and the Drain/Gate DUT terminals are physically switched to the measurement paths. This switching procedure ensures that the time gap while changing to the measurement OM will remain minimum and equal for all DUTs involved in the test. The current is measured for the requested time, except in an $I–V$ characterization in which a single current value is measured for the current bias voltage.

Fig. 10 illustrates the measurement procedure during any measurement execution. The figure shows how the first measurement points represent only leakage because the DUT is still connected to the stress lines, and how after the leakage current measurement is done, the DUT OM is set to the measurement mode.

For the sake of illustration, when $I_{DS}–V_{GS}$, RTN, and BTI tests are specified for 100 DUTs, a test plan with 2500 lines is generated. Execution of this test plan produces 20 800 elementary test instructions that trigger the same number of GPIB commands, as well as 2300 instructions for the DAQ.

After the set of test experiments specified in the test plan are completed, a secure shutdown protocol is initiated to prevent any electrical damage to the IC. First, a general RESET signal is sent to all unit cells in order to set the secure SB OM for all DUTs; second, all SPA HRSMU output terminals are set as open circuits; third, the PCB biasing voltage set by the power supply is removed, and, finally, the GPIB connections with the power supply and the B1500A SPA are terminated.

VI. EXPERIMENTAL RESULTS

This section will show some illustrative experimental results obtained for pristine pMOS transistors using the proposed hardware setup together with the control software and the use of the ADPSPM parallelization technique.

Fig. 11 shows a complete TZV $I_{DS}–V_{GS}$ test executed over 784 devices involving eight different geometries. Each $I_{DS}–V_{GS}$ curve shown in Fig. 11 has been obtained by sweeping the $|V_{GS}|$ bias voltage from 10 mV to 1.2 V while maintaining the $|V_{DS}|$ voltage at 100 mV. The results show a different current level and a large TZV for each geometry.

Fig. 12 shows several measured RTN traces where each device shows unique RTN behavior. For these RTN measurements, the test plan has been created with $|V_{GS}| = 0.5$ V.
and $|V_{DS}| = 0.1$ V with a measurement time of 100 s and a sampling rate of 500 samples per second for each device.

Fig. 13 shows some of the measurement traces extracted from a BTI characterization test executed using the ADPSPM parallelization technique involving 100 pMOS DUTs. The stress voltages applied are $|V_{GS}| = 2.5$ V and $|V_{DS}| = 0$ V while the measurement voltages are set to $|V_{GS}| = 0.6$ V and $|V_{DS}| = 0.1$ V. The measurements start 2 ms after the stress is removed and 500 samples per second were acquired. The presented traces show discrete current steps as a consequence of charge detrapping from traps during the measurement time after the application of overvoltage stress. Thanks to the use of the ADPSPM technique, the total BTI test time is only ~14 h, whereas it would take ~13.5 days (i.e., ~23 times slower) with the conventional serial test approach.

VII. CONCLUSION

This paper describes a complete measurement setup for the characterization of CMOS transistors with array-based ICs, and the design of a full-custom user-friendly software, named TARS, for the automated generation and execution of variability tests. The system has been used to characterize the 65-nm transistors of the ENDURANCE IC, the first of its kind to enable TZV and pipelined TDV characterization, including HCI. The measurement setup enables several tests: ramped voltages ($I_{DS}$–$V_{GS}$, $I_{DS}$–$V_{DS}$), RTN measurements, and BTI and HCI aging evaluations. These tests can be defined in a matter of seconds and thousands of commands are automatically generated for instrumentation and chip control. The described measurement setup in combination with the ENDURANCE IC, empowered by the ADPSPM algorithm, allows the application of smart stress parallelization techniques to an unlimited number of devices. This testing technique smartly overlaps the stress phases of the aging tests, reducing the measurement time from years to days.

ACKNOWLEDGMENTS

The authors would like to thank M. A. Lagos for his support in the design of the PCB.

REFERENCES

[1] K. J. Kuhn et al., “Process technology variation,” IEEE Trans. Electron Devices, vol. 58, no. 8, pp. 2197–2208, Aug. 2011.
[2] M. I. M. Pelgrom, H. P. Tuinhout, and M. Verheijt, “Transistor matching in analog CMOS applications,” in IEDM Tech. Dig., Dec. 1998, pp. 915–918.
[3] S. K. Saha, “Modeling process variability in scaled CMOS technology,” IEEE Des. Test. Comput., vol. 27, no. 2, pp. 8–16, Mar./Apr. 2010.
[4] X. Wang, A. R. Brown, B. Cheng, and A. Asenov, “Statistical variability and reliability in nanoscale FinFETs,” in IEDM Tech. Dig., Dec. 2011, pp. 5.4.1–5.4.4.
[5] X. Wang et al., “Hierarchical variability-aware compact models of 20 nm bulk CMOS,” in Proc. Int. Conf. Simulation Semiconductor Process Devices (SISPAD), Sep. 2015, pp. 325–328.
[6] S. Mahapatra, Fundamentals of Bias Temperature Instability, Springer Series in Advanced Microelectronics, vol. 52. New Delhi, India: Springer, 2016.
[7] P. Monnig et al., “Implementation of CMOS transistors with array-based TARS,” in Proc. 10th Int. Conf. Rel. Maintenability Saf. (ICRMS), Aug. 2014, pp. 944–947.
[8] T. Grasser, K. Rott, H. Reisinger, M. Waltl, J. Franco, and B. Kaczer, “A unified perspective of RTN and BTI,” in Proc. Int. Rel. Phys. Symp. (IRPS), Jun. 2014, pp. 4A.5.1–4A.5.7.
[9] M. Luo, R. Wang, S. Guo, J. Wang, J. Zou, and R. Huang, “Impacts of random telegraph noise (RTN) on digital circuits,” IEEE Trans. Electron Devices, vol. 62, no. 6, pp. 1725–1732, Jun. 2015.
[10] International Roadmap for Devices and Systems (IRDS). (2018). More Moore. [Online]. Available: https://irds.ieee.org/images/files/pdf/2017/2017IRDS_MM.pdf
[11] A. Rahimi, L. Benini, and R. K. Gupta, “Variability mitigation in nanometer CMOS integrated systems: A survey of techniques from circuits to software,” in Proc. IEEE, vol. 104, no. 7, pp. 1410–1448, Jul. 2016.
[12] J. Martin-Martinez et al., “Probabilistic defect occupancy model for NBTI,” in Proc. Int. Rel. Phys. Symp. (IRPS), Apr. 2011, pp. XT.4.1–XT.4.6.
[13] M. A. Alam, K. Roy, and C. Augustine, “Reliability—and Process-variation aware design of integrated circuits—A broader perspective,” in Proc. Int. Rel. Phys. Symp. (IRPS), Apr. 2011, pp. 4A.1.1–4A.1.11.
[14] K. Agarwal, J. Hayes, and S. Nassif, “Fast characterization of threshold voltage fluctuation in MOS devices,” IEEE Trans. Semicond. Manuf., vol. 21, no. 4, pp. 526–533, Nov. 2008.
[15] P. Weckx et al., “Characterization of time-dependent variability using 32k transistor arrays in an advanced HK/MG technology,” in Proc. Int. Rel. Phys. Symp. (IRPS), Apr. 2015, pp. 3B.1.1–3B.1.6.
[16] M. Simicic et al., “Advanced MOSFET variability and reliability characterization array,” in Proc. IEEE Int. Integr. Rel. Workshop, Mar. 2016, pp. 73–76.
[17] H. Awan, M. Hiromoto, and T. Sato, “Variability in device degradations: Statistical observation of NBTI for 3996 transistors,” in Proc. 44th Eur. Solid State Device Res. Conf. (ESSDERC), Sep. 2014, pp. 218–221.
[18] H. Awan, M. Hiromoto, and T. Sato, “BTIarray: A time-overlapping transistor array for efficient statistical characterization of bias temperature instability,” IEEE Trans. Device Mater. Rel., vol. 14, no. 3, pp. 833–843, Sep. 2014.
[19] C. Schlünder, J. M. Berthold, M. Hoffmann, J.-M. Weigmann, W. Gustin, and H. Reisinger, “A new smart device array structure for statistical investigations of BTI degradation and recovery,” in Proc. Int. Rel. Phys. Symp. (IRPS), Apr. 2011, pp. 2B.6.1–2B.6.5.
[20] M. B. da Silva, B. Kaczer, G. Van der Plas, G. I. Wirth, and G. Groeseneken, “On-chip circuit for massively parallel BTI characterization,” in Proc. IEEE Int. Integr. Rel. Workshop Final Rep., Oct. 2011, pp. 90–93.
[21] E. Bury et al., “Statistical assessment of the full VGV/VD degradation space using dedicated device arrays,” in Proc. IEEE Int. Rel. Phys. Symp. (IRPS), Apr. 2017, pp. 2D-5.1–2D-5.6.
[22] A. Whitcombe, S. Taylor, M. Denham, V. Milovanović, and B. Nikolić, “On-chip I-V variability and random telegraph noise characterization in 28 nm CMOS,” in Proc. 46th Eur. Solid-State Device Res. Conf. (ESSDERC), Sep. 2016, pp. 248–251.
[23] M. Simicic et al., “A fully-integrated method for RTN parameter extraction,” in Proc. Symp. VLSI Technol., Jun. 2017, pp. T132–T133.
[24] J. Diaz-PFortuny et al., “A transistor array chip for the statistical characterization of process variability, RTN and BTI/CHC aging,” in Proc. 14th Int. Conf. Synth. Model. Anal. Simulation Methods Appl. Circuit Design (SMACD), Jun. 2017, pp. 1–4.
[27] J. Diaz-Fortuny et al., “A versatile CMOS transistor array IC for the statistical characterization of time-zero variability, RTN, BTI, and HCI,” IEEE J. Solid-State Circuits, vol. 54, no. 2, pp. 476–488, Feb. 2019.

[28] N. Tega et al., “Impact of threshold voltage fluctuation due to random telegraph noise on scaled-down SRAM,” in Proc. IEEE Int. Rel. Phys. Symp. (IRPS), Apr./May 2008, pp. 541–546.

[29] A. Kochanowska, J. Cichosz, and A. Szweczyk, “A new method for RTS noise of semiconductor devices identification,” IEEE Trans. Instrum. Meas., vol. 57, no. 6, pp. 1199–1206, June 2008.

[30] F. M. Puglisi et al., “Guidelines for a reliable analysis of random telegraph noise in electronic devices,” IEEE Trans. Instrum. Meas., vol. 65, no. 6, pp. 1435–1442, Jun. 2016.

[31] B. Kaczer et al., “Origins and implications of increased channel hot carrier variability in nFinFETs,” in Proc. IEEE Int. Rel. Phys. Symp. (IRPS), Apr. 2015, pp. B.5.1–B.5.6.

[32] P. Weckx et al., “Defect-based compact modeling for RTN and BTI variability,” in Proc. IEEE Int. Rel. Phys. Symp. (IRPS), Apr. 2017, pp. CR-7.1–CR-7.6.

[33] V. M. van Santen et al., “Weighted time lag plot defect parameter extraction and GPU-based BTI modeling for BTI variability,” in Proc. IEEE Int. Rel. Phys. Symp. (IRPS), Burlingame, CA, USA, Mar. 2018, pp. P-CR-6.1–P-CR-6.6.

[34] P. Saraza-Cañanca, J. Diaz-Fortuny, R. Castro-Lopez, E. Roca, J. Martin-Martinez, R. Rodriguez, M. Nafría, F. V. Fernandez, “New method for the automated massive characterization of bias temperature instability in CMOS transistors,” in Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE), Florence, Italy, Mar. 2019, pp. 150–155.

[35] W. Wang, V. Reddy, A. T. Krishnan, R. Vattikonda, S. Krishnan, and Y. Cao, “Compact modeling and simulation of circuit reliability for 65-nm/45-nm CMOS technology,” IEEE Trans. Device Mater. Rel., vol. 7, no. 4, pp. 509–517, Dec. 2007.

[36] A. Toro-Frias et al., “Reliability simulation for analog ICs: Goals, solutions, and challenges,” Integration, vol. 55, pp. 341–348, Sep. 2016.

[37] P. Martin-Lloret et al., “A size-adaptive time-step algorithm for accurate simulation of aging in analog ICs,” in Proc. IEEE Int. Symp. Circuits and Systems (ISCAS), May 2017, pp. 1–4.

[38] A. Rahimi, D. Cesaroni, A. Marongiu, R. K. Gupta, and L. Benini, “Improving resilience to timing errors by exposing variability effects to software in tightly-coupled processor clusters,” IEEE Trans. Emerg. Sel. Topics Circuits Syst., vol. 4, no. 2, pp. 216–229, Jun. 2014.

[39] T. Grasser, P. J. Wagner, P. Hehenberger, W. Goes, and B. Kaczer, “A rigorous study of measurement techniques for negative bias temperature instability,” IEEE Trans. Device Mater. Rel., vol. 8, no. 3, pp. 526–535, Sep. 2008.

[40] S. Mahapatra, N. Goel, A. Chaudhary, K. Joshi, and S. Mukhopadhyay, “Characterization methods for BTI degradation and associated gate insulator defects,” in Fundamentals of Bias Temperature Instability, Springer Series in Advanced Microelectronics, vol. 52. New Delhi, india: Springer, 2016, pp. 43–92.

[41] B. P. Linder and T. Ando, “Combined ramp voltage stress and constant voltage stress for optimal BTI voltage acceleration and lifetime modeling,” in Proc. Int. Rel. Phys. Symp. (IRPS), Jun. 2014, pp. XT.7.1–XT.7.4.

[42] Keysight Technologies. (2019), The Parametric Measurement Handbook [Online]. Available: http://literature.cdn.keysight.com/litweb/pdf/5992-2508EN.pdf

[43] Keysight Technologies. (2018), Keysight Technologies B1500A/B1505A/B1506A/B1507A Device Analyzer Series Programming Guide [Online]. Available: http://literature.cdn.keysight.com/litweb/pdf/B1500-90010.pdf

[44] Keysight Technologies. (2018), Keysight E3631A Triple Output DC Power Supply User’s Guide [Online]. Available: http://literature.cdn.keysight.com/litweb/pdf/E3631-90002.pdf

[45] T-2650BV Precision Temperature Forcing System User’s and Maintenance Manual, Thermonics, Mansfield, MA, USA, 2007.

[46] National Instruments. (2018). NI USB-6501 User Guide [Online]. Available: http://www.ni.com/pdf/manuals/757267a.pdf

[47] J. Diaz-Fortuny et al., “TARS: A toolbox for statistical reliability modeling of CMOS devices,” in Proc. IEEE Int. Conf. Synthesis, Modeling, Anal. Simul. Methods Appl. Circuit Design, Jun. 2017, pp. 1–4.

[48] V. Putcha et al., “Smart-array for pipelined BTI characterization,” in Proc. IEEE Int. Integr. Rel. Workshop (IRW), Oct. 2015, pp. 95–98.

[49] V. Putcha, E. Bury, P. Weckx, J. Franco, B. Kaczer, and G. Groeseneken, “Design and simulation of on-chip circuits for parallel characterization of ultrascaled transistors for BTI reliability,” in Proc. IEEE Int. Integr. Rel. Workshop Final Rep. (IRW), Oct. 2014, pp. 99–102.
Javier Martín-Martínez received the M.S. degree in physics from the Universidad de Zaragoza, Zaragoza, Spain, in 2004, and the Ph.D. degree from the Universitat Autònoma de Barcelona (UAB), Barcelona, Spain, in 2009. He was with the Università degli Studi di Padova, Padova, Italy, and IMEC, Leuven, Belgium. He is currently an Associate Professor with UAB. His current research interests include the characterization and modeling of failure mechanisms in MOSFETs and also RRAM characterization and modeling for neuromorphic applications.

Rosana Rodríguez received the Ph.D. degree in electrical engineering from the Universitat Autònoma de Barcelona (UAB), Barcelona, Spain, in 2000. She was with the IBM Thomas J. Watson Research Center, NY, USA, where she was involved in devices and circuits reliability, funded by the Fulbright Program. She is currently an Associate Professor with the UAB. Her current research interests include the electrical characterization and reliability of CMOS devices, and the effect of failures such as dielectric breakdown, BTI and HCI on devices and circuits performance, as well as the process/time-related variability, and study of the resistive switching phenomenon and its applications.

Francisco V. Fernández received the Ph. D. degree in microelectronics from the Universidad de Sevilla, Sevilla, Spain, in 1992. In 1993, he joined Katholieke Universiteit Leuven, Leuven, Belgium, as a Post-Doctoral Research Fellow. From 1995 to 2009, he was an Associate Professor with the Department of Electronics and Electromagnetism, Universidad de Sevilla, where he was promoted to a Full Professor in 2009. He is also a Department Head with IMSE-CNM, Universidad de Sevilla and CSIC. He has authored more than 250 papers in international journals and conferences. His current research interests include microelectronic reliability, and design and design methodologies of analog, and mixed-signal and RF circuits. Dr. Fernández was the Editor-in-Chief of Integration and the VLSI Journal (Elsevier) from 2005 to 2015.

Montserrat Nafria is a Full Professor with the Electronic Engineering Department, Universitat Autònoma de Barcelona, Barcelona, Spain, where she is currently involving in the characterization and modeling of the time-dependent variability of advanced MOS devices, to develop models for circuit reliability simulators. She is also interested in Resistive RAM and graphene-based devices. She has coauthored more than 250 research papers in scientific journals and conferences in these fields.