A Versatile CMOS Transistor Array IC for the Statistical Characterization of Time-Zero Variability, RTN, BTI and HCI

Javier Diaz-Fortuny, Javier Martin-Martinez, Rosana Rodriguez, Rafael Castro-Lopez, Elisenda Roca, Xavier Aragones, Enrique Barajas, Diego Mateo, Francisco V. Fernandez and Montserrat Nafria

***Abstract*—Statistical characterization of CMOS transistor variability phenomena in modern nanometer technologies is key for accurate end-of-life prediction. This work presents a novel CMOS transistor array chip to statistically characterize the effects of several critical variability sources, such as Time-Zero Variability, Random Telegraph Noise, Bias Temperature Instability and Hot-Carrier Injection. The chip integrates 3,136 MOS transistors of both pMOS and nMOS types, with 8 different sizes. The implemented architecture provides the chip with a high level of versatility, allowing all required tests and attaining the level of accuracy that the characterization of the above-mentioned variability effects requires. Another very important feature of the array is the capability of performing massively parallel aging testing, thus significantly cutting down the time for statistical characterization. The chip has been fabricated in a 1.2-V, 65-nm CMOS technology with a total chip area of 1800 × 1800 μm^2.**

***Index Terms*— Bias temperature Instability (BTI), Negative Bias Temperature Instability (NBTI), Positive Bias Temperature Instability (PBTI), Random Telegraph Noise (RTN), Hot Carrier Injection (HCI), variability, aging, degradation, statistical characterization, CMOS, reliability.**

I. Introduction

Reliability has become a serious concern to analog and digital VLSI circuit designers in modern nanometer scale CMOS technologies. Shifts of key parameters, like threshold voltage (V_{th}) and effective mobility (μ_{eff}), at the fabrication process, as well as their progressive variation over the years, are major culprits behind fatal performance deviations in digital and analog integrated circuits [1][2].

One possible way to classify the sources of variability considers whether the variation is dependent on the fabrication process, or dependent on time. Accordingly, they can be classified into Time-Zero (TZV) and Time-Dependent Variability (TDV), which may occur simultaneously. TZV, typically known as spatial or process variability, is a well-known variability source that consists of a constant, either random or systematic, permanent shift of some device parameters (and, thus, a permanent deviation of the nominal circuit performance). This is due to imperfections during the fabrication process, causing effects that worsen with technology scaling like random dopant fluctuations, line edge roughness or gradient effects [3][4]. TDV, on the other hand, includes transient effects, like Random Telegraph Noise (RTN), and aging effects, like Hot-Carrier Injection (HCI) [5][6] and both types of Bias Temperature Instability (BTI) [7][8]: Negative BTI (NBTI) and Positive BTI (PBTI).

The RTN phenomenon in MOSFETs causes random fluctuations between two or more drain current levels due to the stochastic charge/discharge of oxide and interface traps. RTN effects alter normal circuit operation, leading to circuit performance degradation or to performance failure, e.g., failure of SRAM cells or jitter in ring oscillators [9][10]. On the other hand, BTI and HCI result in a gradual shift of transistor parameters over time, e.g., an increase in the absolute value of the threshold voltage (V_{th}) or a decrease of the effective mobility (μ_{eff}), and the magnitude of these variations is strongly related to the device biasing and temperature conditions [11]. Aging, due to BTI and HCI phenomena, has thus become a major concern for long-term circuit functionality.

Providing an accurate and trustworthy characterization of all these TZV and TDV effects in modern CMOS technologies has, therefore, become a key step in the path towards attaining truly reliable integrated circuits (ICs). Since it is not practical to characterize transistors over years, the typical aging characterization procedure uses accelerated aging tests, in which temperature and/or the drain voltage and/or the gate voltage are raised above their nominal values over a much shorter period of time, i.e., the stress time. These high voltages and temperatures are referred to as stress conditions. For instance, elevated voltages are applied to the devices during several periods of stress (whose duration typically increases exponentially and ranges from seconds to hours), followed by current measurement at low voltages to evaluate the impact of the stress on the device performances. This testing procedure is usually known as stress-measurement (SM) cycle [12], and

This work has been supported in part by the TEC2013-45638-C3-R and TEC2016-75151-C3-R Projects (funded by the Spanish MINECO and ERDF) and by the P12-TIC-1481 Project (funded by Junta de Andalucía). Javier Diaz Fortuny acknowledges MINECO for supporting his research activity through the FPI grant BES-2014-067855.

J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez and M. Nafria are with the Electronic Engineering Department (REDEC) group. Universitat Autònoma de Barcelona (UAB), Barcelona 08193, Spain (email: {javier.diaz; javier.martin.martinez; rosana.rodriguez; montse.nafria}@uab.es).

R. Castro-Lopez, E. Roca and F. V. Fernandez are with the Instituto de Microelectrónica de Sevilla, IMSE-CNM, CSIC and Universidad de Sevilla, Spain (email: {castro; eli; francisco.fernandez}@imse-cnm.csic.es).

X. Aragones, E. Barajas, D. Mateo are with the Department of Enginyeria Electrònica, Universitat Politècnica de Catalunya (UPC), Edifici C4, 08034, Barcelona Spain (email: {enrique.barajas; xavier.aragones; diego.mateo}@upc.edu).

allows extracting the main parameters of the transistors and compare them with their pre-stress values, to compute their shifts. During the measurement phase, in some cases, i.e., BTI, the electrical parameters, e.g., the threshold voltage, of the stressed devices start recovering towards their original values immediately after the removal of the stress [13]. Therefore, accurate timing must be imposed in order to get reliable results. Physical models allow extrapolating the results obtained under accelerated test conditions to normal operation conditions [14].

It is worth emphasizing that, independently of dealing with TZV or TDV, and due to the stochastic nature of these phenomena, a large number of devices must be characterized to obtain trustworthy characterization results. Typically, device characterization techniques are conducted using probe stations for on-wafer device measurements. This characterization procedure, which implies physical contact on usually one device under test (DUT) at a time, results in long serial aging test times when thousands of transistors are involved. Additionally, the area required for probe station measurements of a large number of transistors is very large due to the need of including pads for accessing the terminals of each individual transistor [15].

To perform fast and trustworthy statistical characterization of TZV/TDV effects, array structures with thousands of transistors can be used. Then, automatic characterization of thousands of DUTs can be carried out by implementing digital circuitry to control the access to each DUT terminal through the IC pads. This access capability brings another advantage of using array-based ICs: the possibility of parallelizing the aging tests by stressing several DUTs at the same time, which dramatically reduces the overall characterization time. Also, the total area used will be largely reduced, as compared to the probe station approach.

However, designing a transistor array for these types of measurements is not a simple task. As described in Section II, a set of requirements must be fulfilled to carry out a proper characterization of the variability phenomena. Even though several array-based ICs have been reported in the literature, none of them is able to completely and accurately perform statistical characterization of TZV, RTN, BTI and HCI in a single IC chip. Moreover, not all of these reported ICs use the parallelization capability of the array-based approach.

The main objective of this work is to present a versatile DUT array chip, named ENDURANCE, with 3,136 CMOS transistors, which allows performing a trustworthy TZV and TDV characterization, and both serial and parallel stresses of DUTs.

The rest of the paper is organized as follows. Section II presents the design requirements needed to implement an array-based IC for accurate and fast characterization of all variability effects. Section III discusses the state-of-the-art for array-based structures for statistical characterization of TZV and TDV. Section IV describes the internal architecture of the proposed IC. The digital control circuitry and DUT access circuitry are also presented with a description of the chip functionality that endows the high versatility of the ENDURANCE design. Section V describes the procedures that are followed for TZV, RTN, BTI and HCI characterization. In Section VI, experimental results of TZV, RTN, BTI and HCI tests are presented. Finally, conclusions are outlined in section VII.

II. Transistor Array Design For Reliability Characterization: Requirements

As mentioned above, array structures have two advantages for the statistical characterization of TZV and TDV in CMOS transistors: first, a larger number of DUTs can be characterized for a given silicon area, and, second, proper parallelization techniques can be used to significantly speed up the statistical characterization. As for the latter advantage, the use of stress parallelization techniques can be carried out by implementing the SM procedure using the parallel-stress/pipeline-measurement (PSPM) approach [16]. This technique significantly reduces the aging test time compared to a serial implementation of the SM technique, where only one DUT at a time can be characterized, as is usually the case with probe stations. A major constraint for parallelization is that, at any given moment, only one DUT should be under measurement so that the collected data only account for the degradation of that specific DUT. The PSPM technique deals with this constraint by delaying the SM process of each DUT with respect to the previous one, resulting in partially simultaneous (parallel) stress phases and pipeline measurement phases, as illustrated in the example of Fig. 1.

There are two aspects of PSPM techniques that have to be carefully considered for the design of the array. The first one concerns the timing of the different phases that a DUT undergoes. It is important that the timing of all stress and measurement phases are precisely controlled to later perform accurate aging modeling. In addition, it is especially interesting that the elapsed time between stress and measurement phases and the duration of the measurement phases themselves are exactly the same. Otherwise, the data post-processing for statistical modeling would get unnecessarily complex and less information, especially of the recovery phase, would be collected from the DUTs. This implies that either more devices should be measured for the same statistical significance or less reliable statistical models are obtained from the same number of stressed devices. The second aspect concerns the current level when simultaneously stressing a large number of DUTs. This current, typically rising up to mA levels, is normally directed to a common node

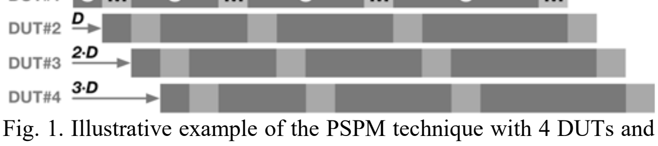

Fig. 1. Illustrative example of the PSPM technique with 4 DUTs and 4 SM cycles, where ‘M’ and ‘S’ stand for measurement and stress phases of the aging test while ‘D’ is the necessary delay to avoid overlapping of the measurement phases and is equal to the duration of a measurement phase. The ‘M’ phases are typically identical while the duration of ‘S’ phases is progressively increased.

(VDD or VSS), and the metal lines carrying it as well as any auxiliary device in the path, e.g., a Transmission Gate (TG), have to be sized and laid out so that adverse effects like electromigration are avoided.

An additional set of requirements emerges in order to enable the characterization of these DUTs under the same conditions as on-wafer characterization. First, when current is flowing through a DUT, a voltage difference appears between the DUT terminals and the externally applied voltage on the corresponding IC pad. This voltage drop is caused by the series resistance of the chip metal paths, the access circuitry and the chip pads. Therefore, for accurate measurements, Force-&-Sense techniques are necessary, meaning that independent Force-&-Sense paths are required to access those DUT terminals where current is flowing. With this access structure, the voltage at the DUT terminal is sensed through the high-impedance Sense path, while the external instrumentation adjusts the Force voltage until the desired voltage value is set in the DUT terminal. Second, calibration techniques are required to compensate leakage currents from the access circuitry that are added to the transistor current being measured. Third, access circuitry, i.e., drain and gate TGs, has to be designed so that it is not degraded by BTI or HCI due to the high voltages applied during the stress periods. A fourth requirement is that digital access and device operation circuitry should be designed to allow all possible variability tests on the DUT: TZV and TDV, as well as supporting parallelized stress of several DUTs with high current flowing through the access circuitry.

Additional must-have features for the DUT array IC are that both nMOS and pMOS transistors of different width/length ratios have to be included or that it must be possible to run variability tests over wide temperature ranges (typically from room temperature to few hundred degrees). The complete list of requirements for complete and trustworthy statistical characterization is summarized as follows:

- *Variety of reliability phenomena to be characterized*: TZV, RTN, BTI and HCI.
- *Variety of DUTs that can be characterized*: the array should include both nMOS and pMOS transistors, with different geometries.
- *Accurate device biasing*: Force-&-Sense on-chip techniques should be available in order to precisely apply the voltages required at the DUT terminals. Furthermore, the design should allow high DC currents during transistor tests.
- *Individual on-chip device access*: the array should allow individual selection of each DUT to separately set its biasing, e.g., to set that DUT on stress or measurement.
- *Reduction of total characterization time*: the array should include the necessary auxiliary circuitry to allow accurately-timed PSPM methods.
- *Robustness of selection circuitry*: all auxiliary circuitry should be designed to avoid its degradation during the application of DUT stress.
- *Accurate current measurement*: the array should provide ways to either calibrate or cancel any leakage current that may distort DUT current measurements.
- *Temperature characterization*: the array should allow device characterization at different temperatures.

III. Previous Works in Array-Based Characterization

Some array-based integrated circuits have been reported for the characterization of TZV, BTI, RTN or HCI reliability effects. Table I summarizes them versus the requirements listed in Section II.

A recent work, which was first used for TZV characterization [17], and later extended to TDV [18] [19], presents a very compact array thanks to a simple unit cell design. This characteristic allows a high device density (32,000 transistors per chip). The array is also able to perform leakage current suppression. Simplicity and compactness are, however, obtained by sharing the drain terminals of all DUTs in each row and the gate terminals in each column. Therefore, the drain Force-&-Sense paths access the array through a line which is shared by all drain terminals of all transistors in a row. Due to the parasitic resistance of this line, the sensed voltage will be different than the Force voltage applied to each device, introducing voltage drop differences that affect parameter extraction. On the other hand, without individual gate and drain TGs, devices in the same column suffer from unwanted stress or, if parallel stress is performed, they are measured at different moments in their recovery phase because only a single device can be measured at the same time. As a consequence, data needs to undergo a complex post-processing process, and less statistical information is collected for a given number of devices compared to an array where stress and measurement times are equal for all devices.

Other arrays also include a large number of DUTs by using simple unit cells designed without TG circuitry [20]-[22]. The array presented in [20] shows a 65-nm test structure including one million modified SRAM cells, which enable individual measurement of pMOS and nMOS transistors for TZV and NBTI characterization. No Force-&-Sense strategy is implemented and the gate terminal is shared by all cells in a row, meaning that after stress, devices are serially measured. This is not a major problem in this work since only the permanent BTI degradation, i.e., after long recovery times, is characterized.

In [22], all DUTs in a row share the gate terminal. Therefore, different portions of the recovery phase will be measured for each DUT, as in [18],[19]. Furthermore, Force-&-Sense connection is implemented as in [17], meaning that voltage drops will exist and the actual DUT biasing is unknown. It includes a leakage reduction structure and, unlike the previously reviewed approaches, it considers different DUT sizes.

The chip presented in [21] contains 96x18 cells, each including 48 DUTs and an A/D converter that serially digitizes the current of each DUT. Due to the selected architecture, no Force-&-Sense scheme is necessary, although a calibration is performed for leakage compensation. However, no stress voltages can be applied, and, hence, only TZV and RTN can be measured.

TABLE I
COMPARISON BETWEEN TRANSISTOR ARRAY CHIPS ACCORDING TO THE MUST-HAVE LIST IN SECTION I

	TECH	**#DEVICES**	**# W/L ratios**	**Supply VDD**		**TZV & I-V**	**RTN**	**NBTI**	**PBTI**	**HCI**	**Force & Sense**	**Temp Range**	**Parallel stress**	**Accurate timing**	**Leakage current**
				CORE	**I/O**										
[17]-[19]	20nm	32,000	1	0.9V	1.8V	YES	YES	YES	YES	NO	Limited	N/A	NO	NO	Partially cancelled
[20]	65nm	1Mbit 6T SRAM	1	N.A.	2.4V	YES	NO	YES	NO	NO	NO	25ºC to 125ºC	NO	NO	Not cancelled
[22]	28nm	54,432	6	1.2V	1.8V	YES	YES	YES	YES	NO	YES	N.A.	NO	NO	Partially cancelled
[21]	28nm	>80,000	2	1V	N.A.	YES	YES	NO	NO	NO	NO	-173ºC to 25ºC	NO	NO	Not cancelled
[24]	180nm	3,996	6	1.8V	2.7V	YES	NO	YES	NO	NO	NO	25ºC to 135ºC	YES	YES	Partially cancelled
[27]	65nm	128	8	1.2V	1.8V	YES	NO	YES	YES	NO	NO	N.A.	YES	YES	Not cancelled
[23]	N.A.	1,300	1	N.A.	N.A.	YES	NO	YES	NO	NO	YES	N.A.	NO	YES	Not cancelled
[26]	28nm	180	1	N.A.	N.A.	YES	NO	YES	YES	NO	YES	N.A.	YES	YES	Not cancelled
[25]	28nm	5,120	1	N.A.	2.1	YES	NO	NO	YES	YES	NO	N.A.	NO	YES	Not cancelled
THIS WORK	**65nm**	**3,136**	**8**	**1.2V**	**Up to 3.3V**	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**	**25ºC to 120ºC**	**YES**	**YES**	**Fully cancelled**

When TGs are included to independently access the terminals of each DUT, the number of devices available in the IC array is necessarily reduced due to the area needed to implement the access circuitry, as in [23]-[27] The use of TGs to control the biasing of each device also enables accurate timing control during the characterization of aging phenomena.

While the array chips proposed in [23] and [26] incorporate on-chip Force-&-Sense paths, those presented in [24] and [27] perform voltage measurement (when a constant current is applied), implying no need for a Force-&-Sense connection, although leakage suppression techniques are used. All four array structures are designed to perform parallel BTI tests, leaving other degradation mechanisms, like HCI, uncharacterized. Similar problems to those described above in [22] are shown in [27], where, after a parallel stress, measurements start at different times for different DUTs.

Arrays including TGs usually incorporate two types of transistors in their design: core transistors for characterization and digital circuitry, and I/O transistors for TGs, avoiding the degradation of TG transistors during stress periods. Their operating voltages, when available, have been included in Table I for better comparison.

While several arrays have been reported to characterize BTI phenomena, characterization of HCI phenomena is reported only in [25]. In this work, however, the characterization cannot be carried out in a parallel fashion. This circuit includes TGs only on the gate terminals (10 gate terminals share a TG), while all drain terminals of one row (with 256 DUTs) are connected. Therefore, voltage drops will appear and this will translate into differences in the drain voltage applied to each DUT. This connection scheme implies that when one DUT is being stressed, other DUTs are also being stressed and, therefore, degraded, making the data analysis much more complex since measured recovery time windows are different for each DUT.

Finally, only [20], [21] and [24] present the characterization of variability phenomena under temperature-controlled conditions.

In the next section, our versatile array IC will be presented. As can be seen in Table I, this transistor array chip fulfills all the requirements listed in section II, being, to the best of the authors' knowledge, the first implementation capable of accurately and statistically characterizing TZV, BTI, HCI and RTN.

IV. THE ENDURANCE CHIP

The architecture of the ENDURANCE IC will be detailed in three different subsections: first, the main building blocks of the chip will be described, then, the unit cell, which corresponds to the basic repeatable IC structure in the array, will be presented and, finally, the operation modes of the chip will be defined.

A. Fundamental building blocks

Fig. 2 shows a photograph of the ENDURANCE chip, which has been fabricated in a 65-nm CMOS technology and encapsulated in a JLCC68 package for testing. The main building blocks of the ENDURANCE IC are shown in Fig. 3. The chip includes 3,136 regular-threshold-voltage MOS transistors (nominally operating at maximum 1.2V) or DUTs distributed over two matrices, one of nMOS transistors and another of pMOS transistors. Each DUT matrix is subdivided into two submatrices of 56 rows and 14 columns, named "Left/Right DUT block", and containing 784 DUTs each. The left DUT blocks contain only DUTs of minimum dimensions, i.e., width W=80nm and length L=60nm, while the right DUT blocks include the 8 different transistor geometries listed in Table II. The total number of devices for each transistor geometry is also included in Table II.

TABLE II
DUT GEOMETRIES DISTRIBUTION IN THE ENDURANCE CHIP

W(nm)	*L*(nm)	# Devices	nMOS Left	nMOS Right	pMOS Left	pMOS Right
80	60	2752	784	592	784	592
200	60	32	0	16	0	16
600	60	32	0	16	0	16
800	60	32	0	16	0	16
1000	60	72	0	36	0	36
1000	100	72	0	36	0	36
1000	500	72	0	36	0	36
1000	1000	72	0	36	0	36

Row and column decoders are used to individually select each DUT from the nMOS or pMOS arrays. Input signals to these decoders are provided by 5-bit shift registers for the column selection, and by 6-bit shift registers for the row selection. Both input serial bit interfaces are accessed through digital I/O pads. Three different bits, named “nMOS/pMOS operation mode bits”, are used to set each DUT terminal biasing in one of the three different operation modes, i.e., stress, measurement and stand-by modes. The combination of the operation modes for each DUT allows the definition of serial tests and the implementation of parallel stress techniques to reduce the total time of the TDV aging tests on hundreds of transistors. During the power-on of the ENDURANCE chip, all DUTs are set to the stand-by mode by means of a general RESET signal, i.e., voltage differences between all DUT terminals are set to 0V.

The DUT terminals are accessed through different physical paths (called “analog signals” in Fig. 3), i.e., there are different paths and pads for drain and gate terminals, which are also independent for each nMOS and pMOS left and right DUT block. Source and bulk terminals are short-circuited and internally connected to VSSA (for nMOS matrices) or VDDA (for pMOS matrices).

In order to allow full variability characterization, each DUT terminal is connected to different analog signal paths depending on the tests to be performed. This is done through different TGs. Each DUT is accompanied by an individual digital circuit that controls 8 TGs, i.e., 5 TGs for the drain and 3 TGs for the gate, which connect the DUT drain and gate to the corresponding on-chip analog paths. Fig. 4 illustrates the distribution of the stress, measure and stand-by analog signal paths for drain and gate DUT terminals, together with the corresponding TGs.

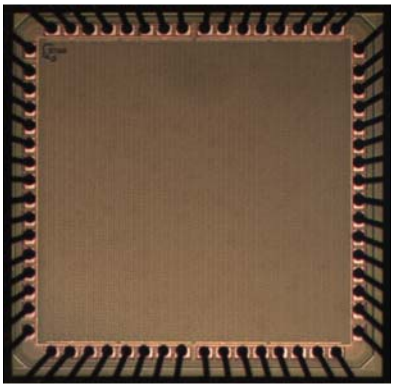

Fig. 2. Photograph of the ENDURANCE chip.

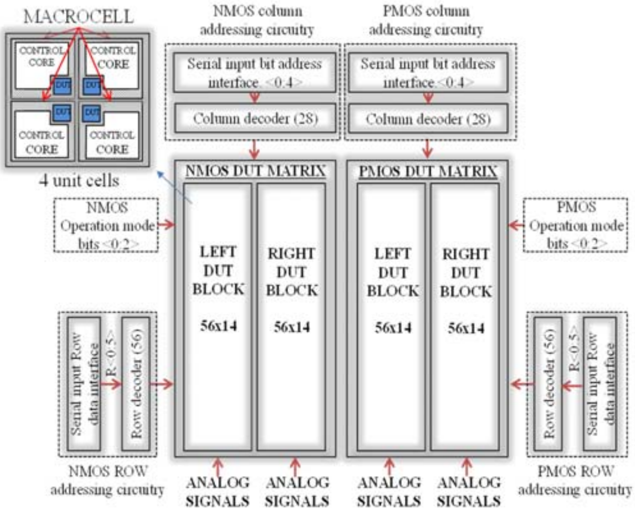

Fig. 3. ENDURANCE top architecture blocks description.

When a DUT is being measured, it is connected to a measure signal path through the corresponding TGs, whereas it is connected to the stress signal paths through dedicated TGs when the DUT is stressed. The circuit design incorporates stand-by signal paths to individually set DUTs into the stand-by mode. Force-&-Sense paths have been designed to access the drain terminal, allowing accurate control of the DUT biasing voltages by minimizing the impact of the IC voltage drop from the IC pads to the DUT drain. The gate terminal does not need a Sense path since the gate current is negligible. For the gate terminal, three different TGs connect the DUT

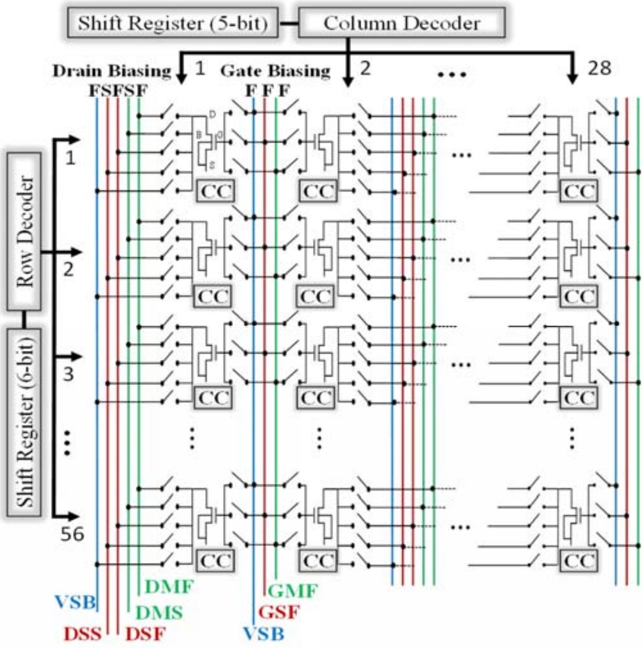

Fig. 4. Schematic diagram illustrating how the DUTs in the arrays are connected to the analog signal paths. In the figure, VSB, DSF, DSS, DMF, DMS, GMF and GSF stand for the DUT matrix stand-by, drain stress force, drain stress sense, drain measure force, drain measure sense, gate measure force and gate stress force paths, respectively. F, S and CC stands for Force and Sense chip paths and the control circuitry, respectively.

terminal to measure Force, stress Force and stand-by Force paths. The connection to the measure, stress or stand-by paths are selected by digital signals that are fed to the individual digital circuitry attached to each DUT, as will be explained in the next section.

The DUT in combination with its control circuitry and the eight TGs constitute the unit cell of the ENDURANCE chip, while the combination of four unit cells defines the macrocell block as can be seen in the inset of Fig. 3.

B. Unit cell circuitry

Fig. 5 shows the details of the unit cell block. It consists of two separate blocks: the digital CONTROL core and the DUT block. The DUT block contains a single nMOS or pMOS transistor with two analog connections: the drain connection and the gate connection.

The CONTROL core of the unit cell involves two different transistor types in order to meet TZV and TDV test specifications: the main digital circuitry is implemented using regular threshold voltage CMOS transistors working between 0V and 1.2V, which correspond to the VDD/VSS biasing voltages in Fig. 5. Unit cell TGs are designed with I/O transistors with a working operation voltage that ranges from 0V to 3.3V, corresponding to the "VDDA/VSSA" voltages in Fig. 5. These I/O transistors are specifically selected to allow the application of the full VDD voltage range, i.e., 0V to 1.2V, during nominal DUT measurements and also allow the possibility of applying stress voltages, up to 3.3V (VDDA), to the DUT terminals during aging tests without suffering from significant degradation themselves. A level shifter block accomplishes the necessary voltage level shift from 1.2V of the digital circuitry to the 3.3V operation voltage level of the I/O transistors of the TGs.

The digital circuitry of the unit cell consists mainly of three single-bit memories, which are implemented with D flip-flops, and are denoted as "FFD" blocks in Fig. 5. Each memory block stores a 1-bit digital signal corresponding to one operation mode. Therefore, three bits are needed to indicate the operation mode: the stand-by bit "XSB", the measure bit "XXM" and the stress bit "XXS", as shown in Fig. 5.

The output signals of the row and column decoders, "ROW" and "COL", are used to select the desired unit cell of the selected array. These signals activate the "ENABLE" signal, which is provided to the three memory blocks.

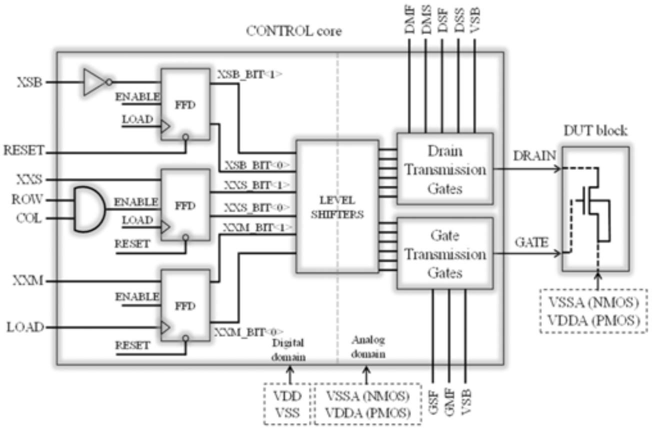

Fig. 5. Unit cell architecture: blocks, interconnections and digital/analog control signals.

When the unit cell has been selected and the operation mode signals have been set, an asynchronous clock pulse has to be sent through the digital line "LOAD" and the operation mode bits are then saved at the flip-flop output until a new operation mode definition is set. A timing diagram illustrating how to set a unit cell in stress mode is shown in Fig. 6.

To reduce the voltage drop from each chip pad to the DUT terminals, all TGs have been designed taking into account the maximum current that will flow through them, and minimizing their resistivity. Special care has been taken in the sizing of the stress TGs connecting the drain terminal of the DUTs since large currents are expected in stress mode. Also, the width of the metal lines connecting all the unit cells has been set to reduce their resistivity and comply with electromigration rules (especially for the analog stress signal lines). On the other hand, the macrocell design ensures minimum distance between the DUTs of the four unit cells which are horizontally and vertically mirrored, as shown in the macrocell zoom in Fig. 3, so that the proximity of the DUTs reduces wafer gradient while the mirroring allows the study of devices with different current orientation. Fig. 7 shows the final layout of the macrocell.

The analog I/O signals of each unit cell for drain and gate biasing and current measurement are labelled with "D" or "G" to specify drain and gate DUT terminals, with "S" or "M" to specify if the paths are for stress or measurement operation modes and with "F" or "S" to distinguish between Force and Sense signals paths. Fig. 5 shows the unit cell DSF/DSS and DMF/DMS path connections for drain stress and measure respectively, GSF/GMF for gate stress and measure respectively and VSB, i.e., voltage stand-by, for both gate and drain DUT terminal connections.

C. Operation modes

Depending on the operation mode defined in the unit cell, each DUT will be connected to the stress, measure or stand-by internal signal paths. In this section, the functionality of these three operations modes, together with an additional mode, the off mode, implemented to improve the versatility of the chip, is discussed.

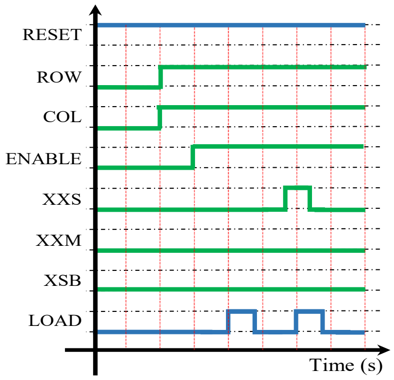

Fig. 6. Timing diagram illustrating how to set a unit cell in stress mode.

1) *Measure mode*

The main objective of this mode is to measure the DUT, e.g., I-V characteristics, in the normal range of voltage biasing, i.e., 0V to 1.2V. This operation mode connects the drain and gate terminals of the DUT to the measure analog signal paths. In the case of the drain terminal, it is connected to the drain measure Force (DMF) and drain measure Sense (DMS) analog signal paths, whereas the gate terminal is connected to the gate measure Force (GMF) analog signal path, as can be seen in Fig. 8 (a). The TG for the DMF signal path has been designed with a suitable W/L ratio to sustain DC currents up to 1.5mA and nominal voltage application, i.e., between 0V and 1.2V.

2) *Stress mode*

This operation mode has been designed to conduct the stress phases of the BTI and HCI aging tests, which consist in the application of an overvoltage to the DUT terminals, i.e., up to 3.3V through both, the drain and the gate, analog stress signal paths. The drain terminal is connected to the drain stress Force (DSF) and drain stress sense (DSS) analog signal paths, whereas the gate terminal is connected to the gate stress Force (GSF) analog signal path, as shown in Fig. 8 (b). The TG for the DSF signal path has been designed with a suitable W/L ratio to sustain DC currents up to 10mA and overvoltage application, i.e., between 0V and 3.3V.

3) *Stand-by mode*

This operation mode connects both drain and gate terminals to the stand-by analog signal paths (VSB) using a single TG, as shown in Fig. 8 (c). This operation mode has been designed to prevent the DUT from being biased, i.e., stressed, when not selected. For the pMOS transistors, the drain and gate terminals are set to VDDA, i.e., 3.3V, resulting in $V_{GS} = V_{DS} = 0V$. In a similar way, for the nMOS transistors, the voltage at the drain and gate terminals is set to VSSA, i.e., 0V, resulting in $V_{GS} = V_{DS} = 0V$.

4) *Off mode*

This operation mode has been designed as a secure operation mode that opens all TGs of all operation modes leaving the DUT disconnected from all analog signal paths. The off mode is used to prevent short circuits between the stand-by mode and other operation modes or to disconnect non-functional DUTs during testing. The off mode can be easily set to any unit cell by changing all 3 operation mode bits to zero. Fig. 6 shows that the off mode is established by setting XXM = XXS = XSB = 0 and sending a load pulse to the unit cell digital circuitry before setting the stress mode.

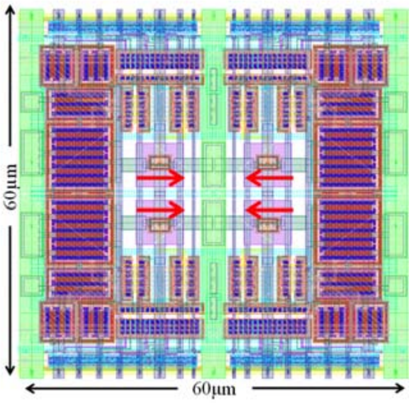

Fig. 7. Layout of an ENDURANCE chip macrocell (60µm x 60 µm). Red arrows distinguish the DUT drain to source current flow direction for each unit cell.

In summary, the ENDURANCE chip design presented in this work fulfills all variability testing requirements listed in Section II. The IC circuit design ensures the ability to conduct trustworthy characterization of all major reliability effects, i.e., TZV, RTN, BTI and, for the first time, HCI aging. The design of the unit cell reduces available area for DUT replication and ensures the control of the biasing state of each DUT terminal at any time. The Force-&-Sense voltage biasing system ensures that, during variability characterization, the IR-drop is mitigated and all voltages defined during testing are correctly applied to the DUT terminals. In addition, the IC utilizes I/O TGs that allow applying stress voltages to the on-chip DUTs up to 3.3V during aging stress. The IC chip incorporates a parallelization technique of the stress phases with accurate synchronization with the measurement phases and leakage current cancellation for each transistor measurement.

V. Time-zero and Time-dependent Variability Phenomena Characterization

In this section, the procedures needed to accurately measure variability phenomena with the ENDURANCE chip will be defined, establishing the requirements for their statistical characterization.

A. *Time-zero variability*

For time-zero variability characterization, the drain current of the DUT, I_{DS}, has to be measured as a function of the gate-source voltage (I_{DS}-V_{GS} curve) and the drain-source voltage (I_{DS}-V_{DS} curve) before any kind of stress is applied. From these measurements, transistor parameters can be extracted, e.g., threshold voltage, V_{th}. Then, a constant voltage is applied

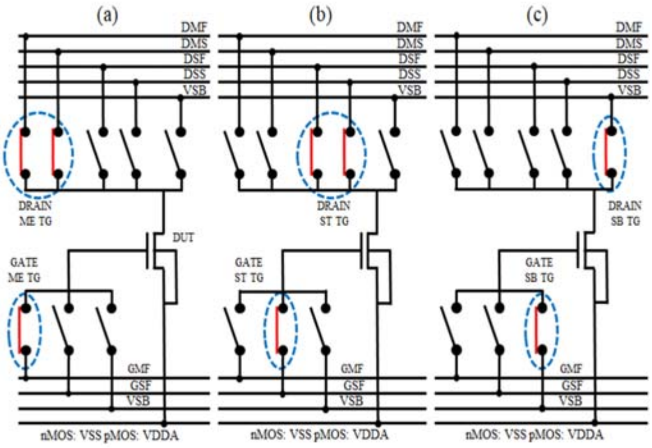

Fig. 8. Schematic diagram of the TG connections that establish the different operation modes: (a) Measure mode (ME); (b) Stress mode (ST); (c) Standby mode (SB).

to the drain terminal |V$_{DS}$| ≤ 100mV while the gate voltage is swept 0V ≤ |V$_{GS}$| ≤ 1.2V for I$_{DS}$-V$_{GS}$ curve measurement. For I$_{DS}$-V$_{DS}$ curve extraction, a staircase sweep is applied to the gate terminal, e.g., |V$_{GS}$| = 0.1V, 0.2V, 0.4V, 0.6V, 0.8V, 1.0V and 1.2V, while, for each |V$_{GS}$| voltage, the drain voltage is swept: 0 ≤ |V$_{DS}$| ≤ 1.2V.

For these measurements, DUTs, which are initially set in a stand-by mode, are serially selected and connected in measure mode. After each DUT has been characterized, it is set back to stand-by mode and a new DUT is selected for characterization.

B. *Transient effects*

The RTN testing procedure consists in the application of constant voltages to the drain and gate DUT terminals while the drain current of the DUT is measured. For the RTN measurements, the voltages are kept at 0V ≤ |V$_{GS}$| ≤ 1.2V and 0V ≤ |V$_{DS}$| ≤ 1.2V.

In this kind of tests, one single DUT is selected and configured in measure mode, while the rest are set to stand-by mode, in order to capture only the drain current fluctuations due to the RTN phenomenon of the selected device.

C. *Time-dependent variability*

The TDV characterization is done in a three-step sequence, where the second and third steps are repeated *M* times to form the SM cycles of aging test patterns. The first step is the I$_{DS}$-V$_{GS}$ initial characterization of the "fresh" DUTs, keeping the voltages in the nominal operating ranges. The second step is a stress phase, where DUTs are aged or degraded by applying larger-than-nominal operating voltages. The third step is a final measurement phase where the drain current is measured at a high sampling rate to assess the degradation of key parameters like the threshold voltage.

The initial characterization is done by setting the DUT in measure mode, with voltage values in the drain and gate terminals of the DUTs ranging from 0V to 1.2V. For the BTI stress phase, DUTs are set into stress mode and an overvoltage is applied for a certain period of time to the gate terminal, e.g., 0V ≤ |V$_{GS}$| ≤ 3.3V, while V$_{DS}$ is kept at 0V. For HCI aging tests, |V$_{GS}$| > 0V and non-zero values of |V$_{DS}$| must be defined, i.e., |V$_{DS}$| ≤ 3.3V. Immediately after the stress phase, the DUT is set back to measure mode to capture the drain current during a fixed period, while the DUT is working in the linear region with V$_{GS}$ ≈ V$_{th0}$ and V$_{DS}$ ≤ 100mV. Changes in V$_{th}$ can be easily tracked from the I$_{DS}$ changes [13]. The aging test patterns are applied equally to all devices involved in the aging test, ensuring the same stress and measurement periods and minimum and equal time gaps when changing from stress to measurement phases.

The total test time needed for a serial BTI/HCI aging test, i.e., one-DUT-at-a-time per SM cycle, can be calculated from the following equation:

$$T_{SERIAL} = N\cdot\left[T_{I_{DS}\text{-}V_{GS}} + T_{STRESS} + T_{I_{DS}\text{-}MEASURE}\right] \tag{0}$$

with:

$$T_{STRESS} = \left(\sum_{j=1}^{M} K^{j-1}\cdot t_s\right) \qquad T_{I_{DS}-MEASURE} = \left(M\cdot t_{measure}\right) \tag{1}$$

where N is the number of DUTs, M is the number of SM cycles, $T_{I_{DS}-V_{GS}}$ corresponds to the time used to get the I$_{DS}$-V$_{GS}$ curve, $T_{I_{DS}-MEASURE}$ is the total measurement time, $t_{measure}$ stands for the measurement time after each stress phase, T_{STRESS} is the total stress time, t_s is the stress time in the first SM cycle, and K defines the growth rate of the stress time periods. For instance, let us consider a first example, denoted as *m5ts1*, with $T_{I_{DS}-V_{GS}}$ = 10s, M = 5, $t_{measure}$ = 100s, t_s = 1s, and K =10. The time required for the aging test of a single DUT would be $T_{SERIAL} \approx$ 3.23 hours. If we consider a second example, denoted as *m3ts100*, where two parameters are changed: M = 3 and t_s = 100s, i.e., the first two SM cycles of example *m5ts1* are skipped, T_{SERIAL} is only reduced to 3.17 hours since the skipped cycles are the shortest ones. When serially performing the same aging tests on a large number of devices, e.g., N = 100, the total test time would rise to approximately 13.4 and 13.2 days respectively, which clearly demonstrates the need for a different alternative, as discussed in Section II. In [16], the PSPM technique was first reported. This method, however, has a critical limitation on the maximum number of DUTs that can be tested simultaneously, as illustrated in Fig. 9, where not more than 4 DUTs could be tested without overlapping of the measurement phases. This maximum number of DUTs is given by:

$$N_{MAX} = \left\lfloor \frac{t_s\cdot K}{t_{measure}} \right\rfloor + 1 \tag{2}$$

For a number of devices, N, below the upper limit N_{MAX}, the total test time T_{PSPM} is (the initial I$_{DS}$-V$_{GS}$ characterization time has been neglected):

$$T_{PSPM} = M\cdot t_{measure} + \left(N-1\right)\cdot t_{measure} + \sum_{j=1}^{M} K^{j-1}\cdot t_s \tag{3}$$

For the example *m5ts1*, N_{MAX} = 1, i.e., parallel stress is not possible. However, in the example *m3ts100* up to 11 DUTs can be tested in parallel. Therefore, the PSPM technique becomes inadequate for certain values of the tuple $\{t_{measure}, t_s, K\}$.

In [28] and [29], a "place-and-check" algorithm is introduced that tries to find the necessary delay of the stress/measurement phases in order to avoid any overlap

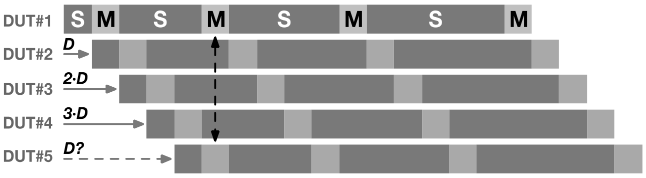

Fig. 9. Illustrative example of the limitation of the PSPM technique.

between any device measurement during the entire test, e.g., DUT#5 in Fig. 9. Again, this non-linear delay depends very much on the values of the tuple $\{ t_{measure},\ t_s,\ K \}$as well as on the number of cycles M. The problem with this approach is that, even though some benefit can be gained when compared with the solution in [16], the most typical outcome is that the DUT aging sequence has to be delayed a complete number of cycles and, essentially, it means that this algorithm repeats the PSPM tests in series until all DUTs are tested depending on the N_{MAX} boundary condition. For the example *m5ts1* above, the "place-and-check" algorithm is unable to reduce the test time with respect to the brute-force serial method since no device can be tested in parallel. However, for the example *m3ts100* the total test time for 100 DUTs is reduced to ~2.48 days since the larger stress time of the second SM cycle $\left(t_s \cdot K\right)$ allows to allocate the measurement of up to 10 other devices during that period of stress.

The solution proposed in this paper uses stand-by periods that can be introduced between certain measurement and stress phases to make the necessary room to accommodate any number of DUTs. The approach is based on the fact that these intermediate stand-by periods will be considered in the data analysis (as an additional recovery time) for the evaluation of the damage suffered by the tested devices under different aging phenomena.

As depicted in Fig. 10, in order to achieve a precise SM parallel test, the algorithm treats and analyses each cycle of the test individually. In the first cycle (C1, in orange in Fig. 10), the SM pattern of each DUT is delayed, like in a regular PSPM approach, in order to ensure that all the measurement phases in all DUTs are pipelined. Once all the SM patterns of cycle C1 have been temporarily allocated, the algorithm starts the parallel distribution of the next cycle (C2). The critical condition is that the last measurement phase of cycle C1 (for DUT#5 in Fig. 10) cannot overlap with the first measurement phase of cycle C2 (for DUT#1). In order to guarantee this unbreakable condition, the algorithm computes the total SM time needed for DUT#1 in cycle C2 (time B), and compares it to the time required for measurement of the remaining DUTs in cycle C1 (time A). If the difference is negative (B<A), the algorithm inserts the necessary standby period (SB) to avoid any measurement overlap, shown in Fig. 10 as green periods. This improved PSPM can thus accommodate any number of DUTs for testing. The total test time of the NEW-PSPM method is:

$$\begin{aligned} T_{NEW-PSPM} &= t_s + M \cdot N \cdot t_{measure} + \\ &+ \sum_{j=2}^{M} \max\left[\left(K^{j-1} \cdot t_s - t_{measure} \cdot \left(N-1\right)\right), 0\right] \end{aligned} \tag{3}$$

For the two examples above, *m5ts1* and *m3ts100,* the total test time of the 100 DUTs using the proposed method reduces to 13.9 and 8.4 hours, respectively, which represents a dramatic improvement with respect to the previous technique. It also demonstrates the ability of the NEW-PSPM method to optimally allocate the stress and measurement periods, independently of the number of devices, number of cycles or stress or measurement times.

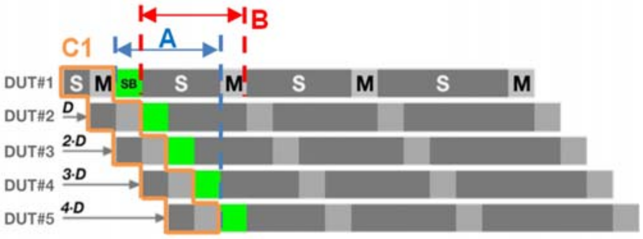

Fig. 10. Illustrative example of the NEW-PSPM technique with 5 DUTs during a 4-cycle SM test. C1 corresponds to the first SM pattern. A denotes the available time after the end of the DUT#1 measurement and B is the needed elapsed time for the SM pattern execution of DUT #1 in cycle C2.

This NEW-PSPM method has been further improved by adding, before each stress phase and for each DUT, an I_{DS}-V_{GS} characterization so that the impact of BTI and HCI effects on the transistor mobility can be obtained. The initial fresh characterization will also be used to extract the fresh V_{th0}, i.e., before any stress has been applied to the DUT, that is used as the applied V_{GS} voltage during the measurement phase.

For a trustworthy characterization of all variability phenomena, the leakage current through the measurement channel is captured before connecting the DUT in the measure mode, and is later used for calibration of the drain current measurements [30].

VI. Experimental Results

For the electrical characterization of the ENDURANCE chip, the setup shown in Fig. 11 (a) has been implemented, which includes the custom-designed Printed Circuit Board (PCB) shown in Fig. 11 (b). The Keysight B1500 semiconductor parameter analyzer (SPA) has been used for voltage biasing of the DUTs using the provided Force-&-Sense connections of its source measurement units (SMU), whereas the Agilent E3631A power supply is used for chip biasing. The temperature of the chip can be controlled with the temperature system Thermonics T-2500E, which allows testing from room temperature to a few hundred degrees.

The interconnection of all the instrumentation equipment has been done by using the standard IEEE 488.1 GPIB BUS, whereas the chip digital control has been accomplished using the USB-6501 digital acquisition system from National Instruments. The test setup is completed with ad-hoc measurement software designed under the Matlab® environment. The software, named TARS [30], allows users to define and monitor variability tests involving thousands of devices within the ENDURANCE chip. The software automatically converts the tests defined by users into all the necessary digital signals for chip control and handles all GPIB functions to control the laboratory instrumentation for chip biasing, current measurement and automatic temperature control.

A. *Time zero variability characterization*

In the present section, a time-zero variability characterization of the DUTs in the ENDURANCE chip is presented. Fig. 12 (a) shows 784 80nm/60nm nMOS I_{DS}-V_{GS} curves, while in Fig. 12 (b) 784 pMOS I_{DS}-V_{GS} curves are

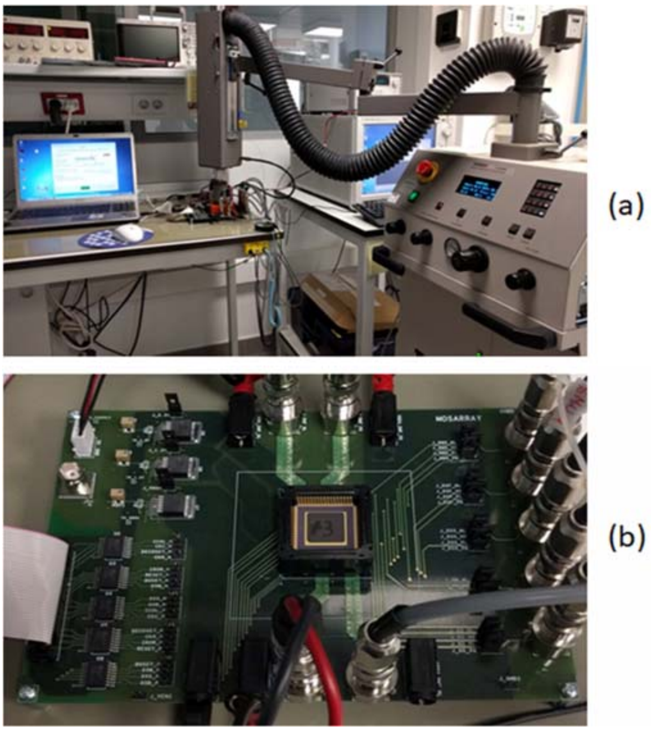

Fig. 11. (a) Laboratory set-up for measurements. (b) Printed circuit board where the ENDURANCE chip is inserted.

shown. Both sets of curves show clear TZV in the fabricated samples. Fig. 12 (c)-(d) shows the row/column matrix distribution of the extracted threshold voltages, V_{th0}, of the 80nm/60nm devices. Fig. 12 (c) shows the V_{th0} distribution across the nMOS DUT Left Submatrix and Fig. 12 (d) shows the V_{th0} distribution of the pMOS DUT Left Submatrix. The values of V_{th0} were obtained by applying the constant current method [31] to each measured I_{DS}-V_{GS} curve. The figure seems to indicate that there is no correlation between the V_{th0} values and the DUT location inside the chip matrices.

B. RTN characterization.

A full RTN characterization has been automatically performed for the 3,136 ENDURANCE chip DUTs. Massive measurements have been conducted serially, DUT by DUT, by means of the TARS control software. Two different tests have been performed for each DUT: an initial I_{DS}-V_{GS} curve extraction (with V_{DS} = 0.1V), to determine the threshold voltage V_{th0} and a constant current measurement of the DUT drain current (I_{DS}) during 100s (with V_{GS} = 0.4V and V_{DS} = 0.1V), that can be mapped to variations of $V_{th.}$

The examples of the RTN results obtained from the ENDURANCE chip, shown in Fig. 13 (a) describe six neat RTN signals converted to the variations of the threshold voltage (ΔV_{th}) as a function of time. The signal plotted in Fig. 13 (b) demonstrates that the system has been capable of capturing I_{DS} changes smaller than 1nA. The inset in Fig. 13 (a) corresponds to the RTN captured signal converted into ΔV_{th} from Fig. 13 (a) showing threshold voltage variations below 10µV [13]. Different charge trapping and detrapping times can be observed in each signal, resulting from different defects emission and capture times. These results confirm that the presented measurement system is a powerful tool to capture the RTN phenomenon using the designed chip.

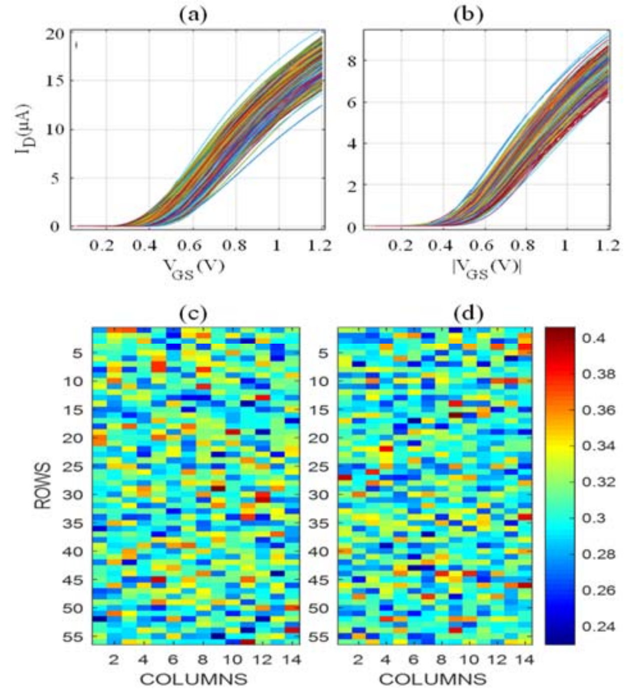

Fig. 12. I_{DS}-V_{GS} curves extracted from the nMOS Left DUT submatrix (a) and the pMOS Left DUT submatrix (b). Initial threshold voltage spatial distribution for the NMOS Left DUT submatrix (c) and for the PMOS Left DUT submatrix (d) for the smallest transistor geometry (80nm/60nm).

C. Bias Temperature Instability characterization.

As an example of the statistical study the BTI phenomena using the chip, 4 BTI tests (each one involving 200 80nm/60nm pMOS devices) have been conducted. Different stress voltages have been applied in each test, i.e., $|V_{GS}|$ = 1.2V, 1.5V, 2V, 2.5V and $|V_{DS}|$ = 0V, while $|V_{GS}|$ = 0.6V and $|V_{DS}|$ = 0.1V have been set to all DUTs in the measurement phase. Devices have been characterized using a 5-cycle SM

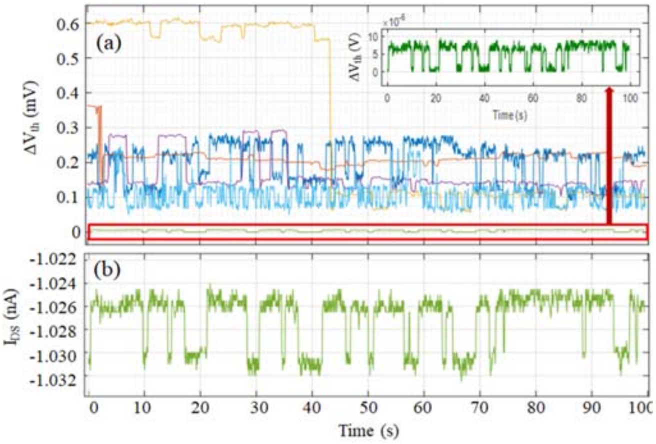

Fig. 13. (a) Six neat RTN signals captured using the ENDURANCE chip during a 3,136 RTN DUT test characterization. The inset in (a) shows a zoom of the ΔV_{th} signal with the smallest amplitude during the test. (b) Original I_{DS} measurement of the signal presented in the inset of Fig. 13 (a).

pattern, in which the duration of the stress phases is increased exponentially, i.e., 1s, 10s, 100s, 1,000s, 10,000s, while all measurement phases (where the drain current was constantly captured) last 100s. The elapsed time between the end of the stress phase and the acquisition of the first drain current value during the measurement phase is 2ms, which corresponds to the maximum sampling rate of the measurement equipment.

Before the execution of the NEW-PSPM technique of the BTI aging test, initial I_{DS}-V_{GS} measurements were performed in order to obtain the equivalent threshold voltage shift ΔV_{th} induced by the BTI test. Thanks to our NEW-PSPM technique to parallelize the 4 BTI tests, the ~107 days of the serial test have been reduced to only ~4.8 days.

In Fig. 14 (a), a set of 13 DUT drain current measurements converted into the resulting ΔV_{th} are shown. A stepwise-recovery trend in the processed ΔV_{th} traces can be clearly recognized in the tested DUTs with small area, corresponding to the discharge of individual traps [32, 33]. Fig. 14 (b) shows the cumulative distribution of ΔV_{th} in a Weibull plot at $t_{measure}$ = 10ms after each stress time. In agreement with [34], the results show that the whole distribution shifts to larger $|\Delta V_{th}|$ values when stress time increases, limiting the device reliability.

D. *Channel Hot Carriers characterization.*

This section presents, as an example, an automatic HCI aging test that involves 200 80nm/60nm nMOS DUTs. The experiment has been divided in four different HCI tests, each one involving 50 DUTs with different V_{GS} and V_{DS} stress voltages combinations: (1) V_{GS} = 1.5V and V_{DS} = 2.4V, (2) V_{GS} = 2.0V and V_{DS} = 2.4V, (3) V_{GS} = 1.5V and V_{DS} = 1.5V and (4) V_{GS} = 2.0V and V_{DS} = 1.5V. These voltage combinations allow users to evaluate and compare the HCI damage when increasing V_{GS} (1.5V to 2.0V) for high and medium V_{DS} overvoltage bias conditions (2.5V and 1.5V). All four HCI aging tests consist in the application to each tested device of 4 SM cycles with I_{DS}-V_{GS} curve characterization before starting any aging measurement and after each SM phase. The duration of the measurement phase, where the current of a single device is measured, is set to 100s, while the time of the stress phase is increased exponentially from 1s to 1000s.

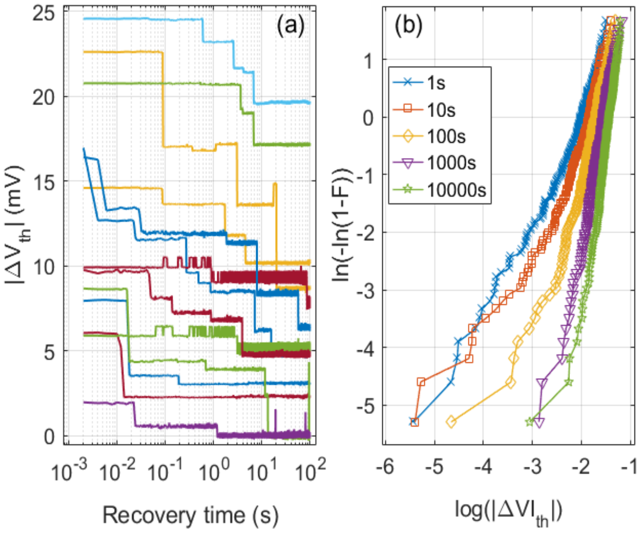

Fig. 14. (a) Typical ΔV_{th} recovery curves obtained during the measurement phases of a BTI test. (b) Cumulative distribution function of ΔV_{th} for 200 pMOS devices at $t_{measure}$ = 10ms after the application of each BTI stress with $|V_{GS}|$ = 2.5V and $|V_{DS}|$ = 0V.

In order to reduce the total HCI test time, the NEW-PSPM parallelization algorithm has been implemented, overlapping the stress times of the DUTs whenever possible. The application of this technique has reduced the time of the 4 HCI tests from ~3.5days, in the case of a serial aging test, to ~22h, which is ~3.7 times faster. To the best of our knowledge, this is the first time that HCI parallel testing is reported. Fig. 15 (a) shows the mean ΔV_{th} as a function of the stress time showing that device degradation is strongly dependent on the V_{GS} and V_{DS} stress conditions [5]. As illustrative example, Fig. 15 (b) shows the cumulative distribution functions (CDF) obtained for the particular case of V_{GS}=2.0V and V_{DS}=2.4V after each stress time (ST).

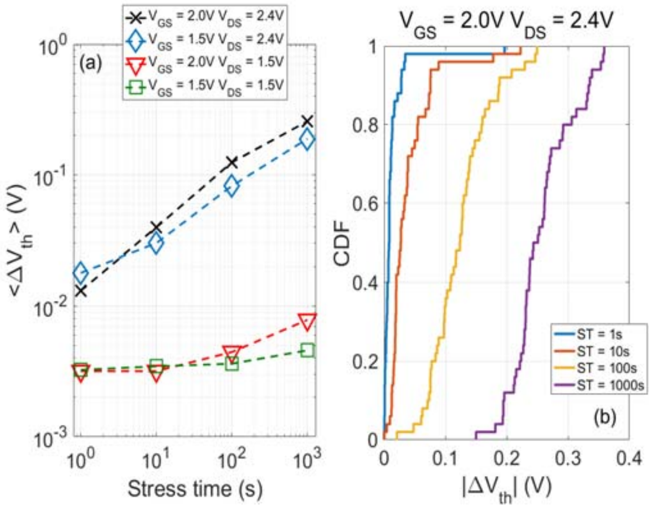

Fig. 15. (a) Mean ΔV_{th} as a function of the stress time for the four HCI tests. (b) CDF plots of the ΔV_{th} values after each ST for the HCI test with V_{GS} = 2.0V and V_{DS} = 2.4V.

VII. Conclusion

In this work, a versatile array chip for the characterization of variability effects in CMOS transistors has been presented. The chip contains sufficient test devices (nMOS and pMOS) of different sizes to get statistically significant results for any test proposed. TZV, RTN, BTI and, for the first time, HCI aging can be evaluated using a unique IC. In this regard, the ENDURANCE chip is the only one of its nature that allows parallel BTI/HCI testing.

The IC incorporates a straightforward circuit design that ensures the ability to perform trustworthy device level reliability characterization. The DUT circuitry added for this purpose, however, limits the number of DUTs that can be included.

The insertion of a Force-&-Sense voltage biasing system into the circuit design guarantees that, during variability characterization, IR-drops are mitigated and all defined voltages are correctly applied to the device terminals. The IC I/O TGs allow applying stress voltages to the DUTs up to 3.3V during aging tests without significant degradation of the access circuitry. In addition, all required variability tests can be performed at different temperatures.

The chip incorporates a novel parallelization technique of the stress phases of aging tests, which ensures accurate synchronization between measurement (with leakage current cancellation) and stress phases, and enables a much faster characterization.

References

[1] J. A. Croon, S. Decoutere, W. Sansen, and H. E. Maes, "Physical modeling and prediction of the matching properties of MOSFETs," in *Proc. 30th Eur. Solid-State Circuits Conf.*, pp. 193–196, 2004.

[2] M. J. M. Pelgrom, H. P. Tuinhout, and M. Vertregt, "Transistor matching in analog CMOS applications," in *Proc. Int. Electron Devices Meet. Tech. Dig.*, pp. 915–918, 1998.

[3] Z. Xie and D. Edwards, "Statistical analysis model of nano-CMOS variability with intra-die correlation due to proximity," in *Proc. 8th EUROSIM Congr. Model. Simul.*, pp. 628–632, Sep. 2013.

[4] A. Asenov, "Simulation of statistical variability in nano MOSFETs," in *Proc. IEEE Symposium on VLSI Technology*, vol. 1, pp. 86–87, 2007.

[5] M. Duan, J. F. Zhang, Z. Ji, W. D. Zhang, B. Kaczer, and A. Asenov, "Key issues and solutions for characterizing hot carrier aging of nanometer scale nMOSFETs," *IEEE Trans. Electron Devices*, vol. 64, no. 6, pp. 2478–2484, 2017.

[6] P. Magnone *et al.*, "Impact of hot carriers on nMOSFET variability in 45- and 65-nm CMOS technologies," *IEEE Trans. Electron Devices*, vol. 58, no. 8, pp. 2347–2353, 2011.

[7] S. Pae, J. Maiz, C. Prasad, and B. Woolery, "Effect of BTI degradation on transistor variability in advanced semiconductor technologies," *IEEE Trans. Device Mater. Reliab.*, vol. 8, no. 3, pp. 519–525, 2008.

[8] A. E. Islam, N. Goel, S. Mahapatra, and M. A. Alam, Fundamentals of Bias Temperature Instability in MOS Transistors, vol. 52. New Delhi: Springer India, 2016.

[9] V. M. van Santen, J. Martin-Martinez, H. Amrouch, M. M. Nafria, and J. Henkel, "Reliability in super- and near-threshold computing: A unified model of RTN, BTI, and PV," *IEEE Trans. Circuits Syst. I*, pp. 1–14, 2017.

[10] F. M. Puglisi, A. Padovani, L. Larcher, and P. Pavan, "Random telegraph noise: Measurement, data analysis, and interpretation," in *Proc. IEEE 24th International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA)*, pp. 1–9, 2017.

[11] T. Grasser, H. Reisinger, P. Wagner, and B. Kaczer, "The time dependent defect spectroscopy (TDDS) technique for the bias temperature instability," in *Proc. Int. Reliab. Phys. Symp. (IRPS)*, vol. 1835, pp. 9–10, 2010.

[12] M. Bhushan and M. B. Ketchen, "Microelectronic Test Structures for CMOS Technology". New York, Springer New York, 2011.

[13] B. Kaczer *et al.*, "Ubiquitous relaxation in BTI stressing-new evaluation and insights," *in Proc. Int. Reliab. Phys. Symp. (IRPS)*, pp. 20–27, 2008.

[14] J. Martin-Martinez *et al.*, "Probabilistic defect occupancy model for NBTI," *in Proc. Int. Reliab. Phys. Symp. (IRPS)*, p. XT.4.1-XT.4.6, 2011.

[15] O. Michael, M. Juettner, and J. Hoentschel, "Mass data analysis of random telegraph noise in 22nm FDSOI back biased transistors," in *Proc. of the EUROSOI-ULIS*, 2017.

[16] T. Sato, T. Kozaki, T. Uezono, H. Tsutsui, and H. Ochi, "A device array for efficient bias-temperature instability measurements," in *Proc. Eur. Solid-State Device Res. Conf.*, pp. 143–146, 2011.

[17] C. S. Chen *et al.*, "A compact test structure for characterizing transistor variability beyond 3σ," *IEEE Trans. Semicond. Manuf.*, vol. 28, no. 3, pp. 329–336, 2015.

[18] P. Weckx *et al.*, "Characterization of time-dependent variability using 32k transistor arrays in an advanced HK/MG technology," *in Proc. Int. Reliab. Phys. Symp. (IRPS)*, p. 3B11-3B16, 2015.

[19] P. Weckx, B. Kaczer, C. Chen, P. Raghavan, D. Linten, and A. Mocuta, "Relaxation of time-dependent NBTI variability and separation from RTN," in *Proc. Int. Reliab. Phys. Symp.*, p. XT9.1-XT9.5, 2017.

[20] T. Fischer, E. Amirante, K. Hofmann, M. Ostermayr, P. Huber, and D. Schmitt-Landsiedel, "A 65nm test structure for the analysis of NBTI induced statistical variation in SRAM transistors," in *Proc. 38th European Solid-State Device Research Conference*, pp. 51–54, 2008.

[21] A. Whitcombe, S. Taylor, M. Denham, V. Milovanovic, and B. Nikolic, "On-chip I-V variability and random telegraph noise characterization in 28 nm CMOS," in *Proc. Eur. Solid-State Device Res. Conf.*, pp. 248–251, 2016.

[22] M. Simicic *et al.*, "Advanced MOSFET variability and reliability characterization array," in *Proc. IEEE Int. Integr. Reliab. Work.*, pp. 73–76, 2016.

[23] 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. Reliab. Phys. Symp. (IRPS)*, pp. 56–60, 2011.

[24] H. Awano, M. Hiromoto, and T. Sato, "Variability in device degradations: Statistical observation of NBTI for 3996 transistors," in *Proc. 44th European Solid State Device Research Conference (ESSDERC)*, pp. 218–221, 2014.

[25] E. Bury *et al.*, "Statistical assessment of the full VG/VD degradation space using dedicated device arrays," *in Proc. Int. Reliab. Phys. Symp. (IRPS)*, p. 2D–5.1–2D–5.6, 2017.

[26] 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. Int. Integr. Reliab. Work. Final Rep.*, pp. 90–93, 2011.

[27] H. Awano, M. Hiromoto, and T. Sato, "BTIarray: A time-overlapping transistor array for efficient statistical characterization of bias temperature instability," *IEEE Trans. Device Mater. Reliab.*, vol. 14, no. 3, pp. 833–843, Sep. 2014.

[28] V. Putcha *et al.*, "Smart-array for pipelined BTI characterization," in *Proc. IEEE International Integrated Reliability Workshop (IIRW)*, pp. 95–98, 2015.

[29] 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 International Integrated Reliability Workshop (IIRW)*, pp. 99–102, 2014.

[30] J. Diaz-Fortuny *et al.*, "TARS: A toolbox for statistical reliability modeling of CMOS devices," in *Proc. 14th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*, 2017.

[31] A. Ortiz-Conde, F. J. García Sánchez, J. J. Liou, A. Cerdeira, M. Estrada, and Y. Yue, "A review of recent MOSFET threshold voltage extraction methods," *Microelectron. Reliab.*, vol. 42, no. 4–5, pp. 583–596, 2002.

[32] T. Grasser *et al.*, "The paradigm shift in understanding the bias temperature instability : from reaction – diffusion to switching oxide traps," *IEEE Trans. Device Mater. Reliab.*, vol. 58, no. 11, pp. 3652–3666, 2011.

[33] B. Kaczer, P. J. Roussel, T. Grasser, and G. Groeseneken, "Statistics of multiple trapped charges in the gate oxide of deeply scaled MOSFET devicesapplication to NBTI," *IEEE Electron Device Lett.*, vol. 31, no. 5, pp. 411–413, 2010.

[34] B. Kaczer *et al.*, "Origin of NBTI variability in deeply scaled pFETs," in *Proc. IEEE Int. Reliab. Phys. Symp. Proc.*, pp. 26–32, 2010.

Javier Diaz Fortuny received his BsC in Technical Telecommunications Engineering with specialisation in electronic systems in 2012, and his MsC in Telecommunications Engineer in 2014 from the Universidad Autònoma de Barcelona (UAB). The same year he joined the group of Reliability Devices and Circuits (REDEC) within the Electronic Engineering Department to pursue the PhD Degree focusing in the characterisation and modelling of time-zero and time-dependent variability in MOSFET devices and array-based integrated circuits.

Javier Martin-Martinez 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), Bellaterra, Spain, in 2009. During his Ph.D. studies, he was with the Università degli Studi di Padova, Padova, Italy, and IMEC, Leuven, Belgium. He is currently Associate Professor at UAB. His main research interests include the characterization and modeling of failure mechanisms in MOSFETs and also RRAM characterization and modeling for neuromorphic applications.

Rosana Rodriguez received the Ph. D. in Electrical Engineering from Universitat Autonoma de Barcelona (UAB) in 2000. Funded by the Fulbright program, she worked on devices and circuits reliability at the IBM Thomas J. Watson Research Center (USA). Currently, she is associate professor at the UAB. Her main research interests are focused on 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. Her current research topics also include the study of the resistive switching phenomenon and its applications.

Rafael Castro-López obtained the PhD in Microelectronics from the University of Seville, Seville, Spain, in 2005. Since 1998, he has been a researcher with the Institute of Microelectronics of Seville (IMSE-CNM), where he holds the position of Tenured Scientist. He has participated as a Researcher in several national and international R&D projects. He has co-authored more than 100 international journals and conferences, and has authored or edited 5 books and book chapters. Dr. Castro has served as General Chair and participated in the Program Committee of several international conferences. He is currently serving as Associate Editor of the Integration, the VLSI Journal (Elsevier), and as expert collaborator in the ICT area of the State Research Agency. His current research interests include the design and design methodologies of analog, mixed-signal, and RF circuits, and reliable circuit design.

Elisenda Roca received the Ph.D. degree in physics from the University of Barcelona, Spain, in 1995. Since 1995, she has been with the Institute of Microelectronics of Seville, (IMSE-CNM-CSIC), Seville, Spain, where she is currently a Tenured Scientist. She has been involved in several national and international research projects with different institutions: CEC, ESA or ONR-NICOP. She has also co-authored more than 100 papers in international journals, books, and conference proceedings. Her research interests include modeling and design methodologies for analog, mixed-signal and RF integrated circuits, and reliability circuit design.

Xavier Aragones received the M.Sc. degree in Telecommunication Engineering and the Ph.D. degree (with honors) in Electronic Engineering from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 1993 and 1997, respectively. Since 1998 he is Associate Professor with the Department of Electronic Engineering at UPC, where he is responsible of several graduate courses related to microelectronic design for analog, mixed-signal and RF. His research interests include effects of variability and aging on RF-ICs, substrate noise coupling and isolation; use of temperature as a built-in observer for RF-ICs; front-ends for RF and mmW communications.

Diego Mateo received the M.Sc. degree in telecommunication engineering and Ph.D. degree in electronic engineering from the Universitat Politecnica de Catalunya (UPC), Barcelona, Spain, in 1993 and 1998, respectively. During 2002, he was with the High-Speed and RF Design Group, Wireless Research Laboratory, Lucent Technologies, Bell Laboratories, Murray Hill, NJ, where he was involved in the design of RF front ends for base stations and on the analysis of the effects of substrate noise on RF blocks. He is currently a Full-Time Associate Professor with the Department of Electronic Engineering, Telecommunication Engineering School, UPC. He has coauthored two books and many international journal and conference papers. He holds twenty patents. His research interests include mixed-signal and RF integrated circuits, RF characterization by thermal monitoring and aging and variability issues in integrated circuits.

Francisco V. Fernández got the Ph. D. degree in Microelectronics from the University of Seville, Spain, in 1992. In 1993, he worked as a postdoctoral research fellow at Katholieke Universiteit Leuven (Belgium). From 1995 to 2009, he was an Associate Professor at the Dept. of Electronics and Electromagnetism of University of Seville, where he was promoted to Full Professor in 2009. He is also a Department Head at IMSE-CNM (University of Seville and CSIC). Dr. Fernández was the Editor-in-Chief of Integration, the VLSI Journal (Elsevier) from 2005 to 2015. His research interests lie in microelectronic reliability, and design and design methodologies of analog, mixed-signal and RF circuits. He has authored or edited 5 books and has co-authored more than 250 papers in international journals and conferences. He has served as General Chair of three international conferences and regularly serves at the Program Committee of several international conferences. He has also participated as researcher or main researcher in numerous national and international R&D projects.

Montserrat Nafría is Full Professor at the Electronic Engineering Department of the Universitat Autonoma de Barcelona. Currently, she is working on the characterization and modelling 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 co-authored more than 250 research papers in scientific journals and conferences in these fields. WoS Researcher ID: J-8231-2014; ORCID ID: 0000-0002-9549-2890