Tesis Doctoral/Doctoral Thesis

Doctorado en Ciencias y Tecnologías Físicas

Study of variability phenomena on CMOS technologies for its mitigation and exploitation

A PhD Dissertation by Pablo Sarazá Canflanca

Advisors:

Francisco V. Fernández Fernández

Rafael Castro López

Date: September 10, 2021

Acknowledgements
Agradecimientos

Son muchas las personas que me han ayudado durante los últimos cuatro años a desarrollar el trabajo que ha desembocado en la escritura de esta tesis. Me gustaría empezar agradeciendo su ayuda a mis directores, Paco y Rafa, y a Eli, que ha sido también mi directora en todo menos en el nombre. Estoy seguro de que sin su ayuda, su paciencia, su generosidad hacia mí, nuestras innumerables discusiones y sus incontables revisiones de mi trabajo esta tesis no existiría. Creo haber aprendido mucho de ellos, y espero que en el futuro nuestros caminos sigan entrelazados.

Quisiera agradecer también su apoyo a todas las personas que me han acompañado durante estos años en el IMSE. Nunca me ha faltado una ayuda desinteresada cuando la he necesitado, y puedo decir que he recibido mucho más de lo que he dado. Espero poder equilibrar la balanza en el futuro, aunque no será fácil. Quisiera agradecer su ayuda a las Unidades Técnicas, en especial a Joaquín y Antonio Ragel, a la Unidad de Administración, a Juanma y Juan por su ayuda en el laboratorio, a Piedad por su ayuda en un campo que era nuevo para todos nosotros, y a los jóvenes veteranos Luis Camuñas y Juan Núñez por su ayuda y amistad. Aunque es imposible nombrar a todo el mundo, gracias también al resto de personas que me han acompañado y ayudado durante estos años, ya sea durante nuestras comidas en cafetería, en los pasillos del IMSE, o por las tardes, cuando el IMSE se vaciaba (gracias, María José, por tu compañía).

Como no, quisiera agradecer también el apoyo técnico continuo y, sobre todo, el cariño personal, de todos mis compeñaros de grupo y de despacho, en especial de Antonio Toro, Fabio, Dani, Héctor, José Fernando, Macarena y, más recientemente, Andrés y Eros. En este grupo incluyo también a Manu, que no ha sido compañero de despacho pero sí vecino, y a Diego, compañero de tantos desayunos. Todos ellos han hecho del IMSE un segundo hogar para mí, y lamento que la situación del último año y medio no nos haya permitido vernos tanto como hubiéramos deseado.

Aunque he realizado mi trabajo en el IMSE, gran parte del mismo se ha desarrollado en colaboración con la UAB. En este sentido, quisiera agradecer su ayuda a Montse, Javi Martín y Rosana. He disfrutado (y creo que aprendido) mucho de nuestras discusiones, y espero que volvamos a colaborar en el futuro.

Proveniente también de la UAB, quisiera hacer una mención especial a Javi Díaz. Primero por toda su ayuda en el laboratorio y por videollamada cuando trabajaba

todavía en la UAB. Después, por ayudarme a realizar una estancia de tres meses en IMEC durante mi doctorado, y, finalmente, por ser en gran parte responsable de mi próxima etapa vital. Quisiera dar las gracias también al resto de personas con las que trabajé en IMEC por su ayuda, y en especial a Erik Bury.

Finalmente, quisiera agradecer a mi familia todo lo que ha hecho por mí para que yo haya podido llegar hasta aquí. No sólo, ni especialmente, durante estos últimos cuatro años, sino durante toda mi vida. Gracias a mi madre, a mi padre, a mi hermano Juan y a mi hermana Amaia.

A mi aituna y mi amona,
y a mis abuelos.

Publications and awards related to this Thesis

International journal papers as first author:

- **P. Saraza-Canflanca**, J. Martin-Martinez, R. Castro-Lopez, E. Roca, R. Rodriguez, M. Nafria and F.V. Fernandez, "A detailed study of the gate/drain voltage dependence of RTN in bulk pMOS transistors", Microelectronic Engineering, vol. 215, 111004, 2019.
- **P. Saraza-Canflanca**, J. Diaz-Fortuny, R. Castro-Lopez, E. Roca, J. Martin-Martinez, R. Rodriguez, M. Nafria and F.V. Fernandez, "A robust and automated methodology for the analysis of Time- Dependent Variability at transistor level" Integration, the VLSI Journal, vol. 72, pp. 13-20, 2020.
- **P. Saraza-Canflanca**, H. Carrasco-Lopez, A. Santana-Andreo, P. Brox, R. Castro-Lopez, E. Roca and F.V. Fernandez, "Improving the reliability of SRAM-based PUFs under varying operation conditions and aging degradation" Microelectronics Reliability, vol. 118, 114049, 2021.
- **P. Saraza-Canflanca**, J. Martin-Martinez, R. Castro-Lopez, E. Roca, R. Rodriguez, F. V. Fernandez and M. Nafria "Statistical characterization of Time-Dependent Variability defects using the Maximum Current Fluctuation", IEEE Transactions on Electron Devices, 2021.

Contributions to conferences as first author:

- **P. Saraza-Canflanca**, J. Diaz-Fortuny, A. Toro-Frias, R. Castro-Lopez, E. Roca, J. Martin-Martinez, R. Rodriguez, M. Nafria and F.V. Fernandez, "Automated massive characterization of RTN using a transistor array chip", in International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2018.
- **P. Saraza-Canflanca**, D. Malagon, F. Passos, A. Toro, J. Nuñez, J. Diaz-Fortuny, R. Castro-Lopez, E. Roca, J. Martin-Martinez, R. Rodriguez, M. Nafria and F.V. Fernandez, "Design considerations of an SRAM array for the statistical validation of time-dependent variability models", in International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2018.

- **P. Saraza-Canflanca**, J. Diaz-Fortuny, R. Castro-Lopez, E. Roca, J. Martin-Martinez, R. Rodriguez, M. Nafria, F.V. Fernandez, "New method for the automated massive characterization of Bias Temperature Instability in CMOS transistors", in Design Automation and Test in Europe (DATE), 2019.
- **P. Saraza-Canflanca**, J. Diaz-Fortuny, R. Castro-Lopez, E. Roca, J. Martin-Martinez, R. Rodriguez, M. Nafria and F.V.Fernandez, "TiDeVa: A toolbox for the automated and robust analysis of Time-Dependent Variability at transistor level", in International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2019.
- **P. Saraza-Canflanca**, H. Carrasco-Lopez, P. Brox, R. Castro-Lopez, E. Roca and F.V.Fernandez, "Improving the reliability of SRAM-based PUFs in the presence of aging", in International Conference on Design & Technology of Integrated Systems in Nanoscale Era (DTIS), 2020.
- **P. Saraza-Canflanca**, H. Carrasco-Lopez, P. Brox, R. Castro-Lopez, E. Roca and F. V. Fernandez, "Improving the reliability of SRAM-based PUFs under varying conditions", in Conference on Design of Circuits and Integrated Systems (DCIS), 2020.
- **P. Saraza-Canflanca**, E. Camacho-Ruiz, R. Castro-Lopez, E. Roca, J. Martin-Martinez, R. Rodriguez, M. Nafria and F.V.Fernandez, "Simulating the impact of Random Telegraph Noise on integrated circuits", in International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2021.

Co-authored international journal papers:

- J. Diaz-Fortuny, **P. Saraza-Canflanca**, R. Castro-Lopez, E. Roca, J. Martin-Martinez, R. Rodriguez, F.V. Fernandez and M. Nafria, “Flexible setup for the measurement of CMOS Time-Dependent Variability with array-based integrated circuits”, IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 3, pp. 853-864, 2019.
- J. Diaz-Fortuny, **P. Saraza-Canflanca**, R. Rodriguez, J. Martin-Martinez, R. Castro-Lopez, E. Roca, F.V. Fernandez and M. Nafria, “Statistical threshold voltage shifts caused by BTI and HCI at nominal and accelerated conditions”, Solid State Electronics, p. 108037, 2021.
- G. Pedreira, J. Martin-Martinez, **P. Saraza-Canflanca**, R. Castro-Lopez, R. Rodriguez, E. Roca, F.V. Fernandez and M. Nafria, “Unified RTN and BTI statistical compact modeling from a defect-centric perspective”, Solid State Electronics, p. 108112, 2021.

Co-authored conference contributions:

- J. Diaz-Fortuny, **P. Saraza-Canflanca**, A. Toro-Frias, R. Castro-Lopez, J. Martin-Martinez, E. Roca, R. Rodriguez, F.V. Fernandez and M. Nafria, "A

model parameter extraction methodology including Time-Dependent Variability for circuit reliability simulation", in International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2018.

- G. Pedreira, J. Martin-Martinez, J. Diaz-Fortuny, **P. Saraza-Canflanca**, R. Rodriguez, R. Castro-Lopez, E. Roca, F.V. Fernandez and M. Nafria., "A new time efficient methodology for the massive characterization of RTN in CMOS devices", in IEEE International Reliability Physics Symposium (IRPS), 2019.

- A. Toro-Frias, **P. Saraza- Canflanca**, F. Passos, P. Martin-Lloret, R. Castro-Lopez, E. Roca, J. Martin-Martinez, R. Rodriguez, M. Nafria and F.V. Fernandez, "Generation of lifetime-aware pareto-optimal fronts using a stochastic reliability simulator", in Design Automation and Test in Europe (DATE), 2019.

- , M. Nafria, J. Diaz-Fortuny, **P. Saraza-Canflanca**, J. Martin-Martinez, E. Roca, R. Castro-Lopez, R. Rodriguez, P. Martin-Lloret, A. Toro-Frias, D. Mateo, E. Barajas, X. Aragones and F. V. Fernandez, "Circuit reliability prediction: challenges and solutions for the device time-dependent variability characterization roadblock", in IEEE Latin America Electron Devices Conference (LAEDC), 2021.

- E. Camacho-Ruiz, **P. Saraza-Canflanca**, R. Castro-Lopez, E. Roca, and F.V.Fernandez, "A study of SRAM PUFs reliability using the Static Noise Margin", in International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2021.

Contributions to book chapters:

- J. Martin-Martinez, J. Diaz-Fortuny, A. Toro-Frias, P. Martin-Lloret, P. Saraza-Canflanca, R. Castro-Lopez, R. Rodriguez, E. Roca, F. V. Fernandez, M. Nafria, "Modeling of variability and reliability in analog circuits", chapter from "Modelling methodologies in analogue Integrated Circuit design".

Awards:

- EDA Competition Winner in the SMACD2019 Conference, held in Lausanne on July 2019, for the work "TiDeVa: a toolbox for the automated and robust analysis of Time-Dependent Variability at transistor level", co-authored by J. Diaz-Fortuny, R. Castro-López, E. Roca, J. Martin-Martinez, R. Rodriguez, M. Nafria and F. V. Fernandez.

- Best Paper Award in the DCIS2020 Conference, held virtually on November 2020, for the work "Improving the reliability of SRAM-based PUFs under varying conditions", co-authored by H. Carrasco-Lopez, P. Brox, R. Castro-Lopez, E. Roca and F. V. Fernandez.

- Best Paper Award Runner-Up in the SMACD2021 Conference, held virtually on July 2021, for the work "Simulating the impact of Random Telegraph Noise on integrated circuits", co-authored by E. Camacho-Ruiz, R. Castro-Lopez, E. Roca, J. Martin-Martinez, R. Rodriguez, M. Nafria and F.V.Fernandez.

List of abbreviations

- **BER:** Bit Error Rate
- **BL:** Bitline
- **BTI:** Bias Temperature Instability
- **CDF:** Cumulative Distribution Function
- **CMOS:** Complementary Metal-Oxide-Semiconductor
- **CRP:** Challenge Response Pair
- **CUT:** Circuit Under Test
- **DRV:** Data Retention Voltage
- **DUT:** Device Under Test
- **ECC:** Error Correcting Code
- **GPIB:** General Purpose Interface Bus
- **HCI:** Hot-Carrier Injection
- **HDA:** Helper Data Algorithm
- **IC:** Integrated Circuit
- **IID:** Independent and Identically-Distributed
- **IoT:** Internet of Things
- **MCF:** Maximum Current Fluctuation
- **MLE:** Maximum Likelihood Estimation
- **MSE:** Mean Squared Error
- **MTSV:** Maximum Trip Supply Voltage
- **NBTI:** Negative Bias Temperature Instability
- **NMOS:** N-channel Metal-Oxide-Semiconductor
- **NVM:** Non-Volatile Memory

- **PBTI:** Positive Bias Temperature Instability
- **PDF:** Probability Density Function
- **PDO:** Probabilistic Occupancy Model
- **PMOS:** P-channel Metal-Oxide-Semiconductor
- **PSO:** Particle Swarm Optimization
- **PUF:** Physical Unclonable Function
- **PV:** Process Variability
- **RAD:** Reliability-Aware Design
- **RFID:** Radio-Frequency Identification
- **RNG:** Random Number Generator
- **RTN:** Random Telegraph Noise
- **SRAM:** Static Random-Access Memory
- **TCAD:** Technology Computer Aided Design
- **TDDB:** Time-Dependent Dielectric Breakdown
- **TDV:** Time-Dependent Variability
- **TRNG:** True Random Number Generator
- **TZV:** Time-Zero Variability
- **WL:** Wordline

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Complementary Metal-Oxide-Semiconductor (CMOS) technology has been present for over half a century now [1]. During this time, it has experienced a continuous dimension scaling that has led to tremendous improvements in integrated circuits (ICs), such as an increase in transistor density or a decrease in power dissipation [2]. However, this has occurred at the expense of a much larger variability of some fundamental transistor parameters, such as their threshold voltage or mobility. This variability can be classified into two major categories: Time-Zero Variability (TZV) and Time-Dependent Variability (TDV).

TZV, also known as Process Variability (PV), denotes deviations of the device parameters from their nominal values that occur during the manufacturing process (i.e., at Time-Zero) [3]. These deviations translate into a variability of the IC performances, and can lead to a yield reduction (i.e., a reduction in the percentage of fabricated circuits that meet the design specifications once that the production process is completed) [4]. In contrast to TZV, TDV refers to deviations in the device parameters that appear with time (i.e., Time-Dependent), during the circuit operation. These deviations can also translate into the degradation of the IC performances, and may eventually lead to a fatal failure of the circuit, thus reducing the time-dependent yield, a concept introduced to mirror that of yield for TZV [5]. A schematic representation of yield and its evolution with time, together with the impact that transistor scaling can have on it, is shown in Fig. 1.1. TDV comprehends several phenomena, which include transient effects such as Random Telegraph Noise (RTN) [6], and aging effects such as Bias Temperature Instability (BTI) [7] or Hot-Carrier Injection (HCI) [8].

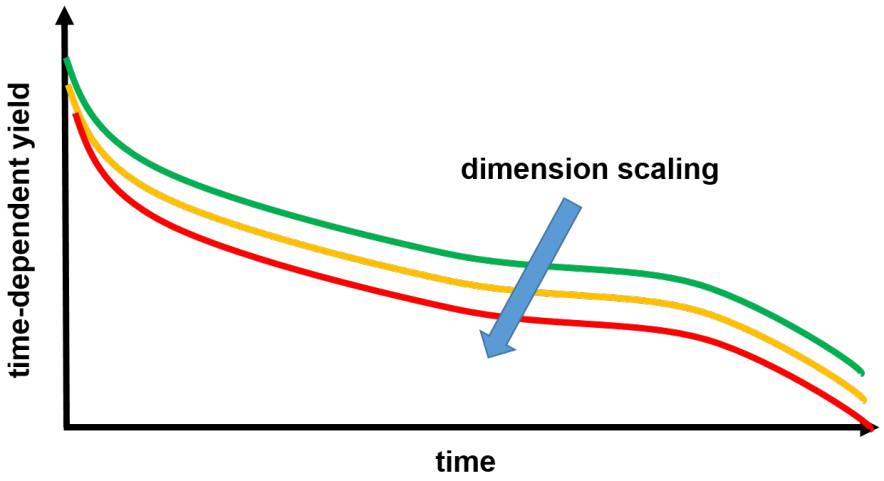

Figure 1.1: Representation of the evolution of yield with time, together with the impact that transistor scaling can have on it.

The work presented in this Thesis focuses on the above-mentioned variability phenomena and their effects on integrated circuits. However, it tackles the variability issue through two very different approaches: the mitigation of its potentially harmful impact on IC reliability, and its exploitation in the field of hardware security. In fact, safety, reliability and security have been considered as a fundamental feature of any modern electronic system in the Strategic Research Agenda for Electronic Components and Systems by the ECSEL Joint Undertaking [9].

Both the mitigation and the exploitation of the impact of TZV and TDV phenomena on circuits, involve a series of steps that are tightly linked to each other. First, the effect of these phenomena must be characterized at device level. Then, the parameters extracted during this characterization step can be used to construct models that describe those phenomena. These models can then be embedded in different simulation tools that allow to predict the impact that variability phenomena will have on circuits under different operation conditions.

In deeply-scaled CMOS technologies (i.e., in the nanometer range), both TZV and TDV phenomena display an stochastic nature. Therefore, their characterization must be massive (i.e., involving hundreds or thousands of devices at different conditions) so that statistically significant information can be extracted. **The first part of this Thesis revolves around such a massive experimental characterization of TDV phenomena, and the consequent construction of a stochastic model for TDV phenomena, the Probabilistic Defect Occupancy (PDO) model** [10]. Such a statistical characterization of TZV is not tackled in this work, since, as it will be seen in the next Section, commercial design tools already allow the assessment of TZV through model files provided by the semiconductor foundries.

Once that a TDV model has been integrated into a simulation tool, it is possible for designers to use such tool to achieve reliability-aware design (RAD) of electronic circuits, i.e., a circuit design process in which the impact of TDV phenomena on the circuit is accounted for [11].

As it has been stated, the second approach to variability in this Thesis is its exploitation to the user's benefit, in particular for security applications. This type of exploitation is performed by Physical Unclonable Functions (PUFs). In the growing field of the Internet of Things (IoT), where an enormous amount of interconnected devices interacts and share information, PUFs have become a very interesting option to secure any sensitive information [12]. PUFs ensure the authenticity of systems through the utilization of TZV, in a manner somehow similar to the identification of human individuals by biometric measurements such as the fingerprint. The basic operation principle of PUFs consists in the application of challenges, and the return from the PUF of responses to those challenges. The response returned after a challenge by a given PUF instance is determined by the stochastic and unpredictable deviations of the parameters of the PUF circuit, and therefore are unique to each PUF instance. It is also important that the response of a given PUF instance to the same challenge remains constant in time to ensure security, which can be challenging when factors such as environmental (i.e., voltage and temperature) variations, or circuit degradation caused by TDV phenomena, are considered. **The second part of this Thesis is focused on the analysis of a new method that aims to improve the reliability of this type of response in SRAM PUFs (i.e., its robustness to TDV-induced degradations), which are one of the most common types of PUFs**.

In the rest of this Introduction chapter, the basic framework is settled to understand the rest of the work presented in the Thesis.

1.2 Time-Zero Variability in CMOS technologies

TZV in CMOS technologies comprises deviations of the transistor parameters from their intended nominal values that occur during the manufacturing process. Although TZV has always been a critical aspect of semiconductor fabrication [13], recent years have brought a growing concern around TZV in deeply-scaled CMOS technologies [14], [15], [16]. TZV can be loosely classified in two categories: intradie, which refers to the changes in the parameters of identical transistors across a short distance, and interdie, which refers to the changes in the parameters of identical transistors across larger distances or fabricated at different times.

TZV must be taken into account during circuit design to mitigate its potentially harmful impact on circuit performances. To this end, the different sources of TZV in advanced CMOS technologies must be considered. Some of the most important ones are random discrete doping, line-edge roughness, interface roughness and oxide thickness variation, polysilicon granularity and high-k dielectric morphology [15]. These can cause variability in a number of transistor parameters, such as as the threshold voltage, the effective source-drain series resistance, the subthreshold current, the device transconductance or the mobility [15], [17], [18]. In turn, this variability at the device level leads to performance variability at the circuit level. Therefore, it is critical to accurately model TZV in order to account for it during the process of reliability-aware design. This can be done through the utilization

of compact MOSFET models, such as the BSIM models, that have been broadly used for the last decades [19]. The evaluation of the impact of TZV through these models can be done, for example, by the utilization of Monte-Carlo or worst-case corner analysis. Nowadays, the corresponding model files are commonly provided by the semiconductor foundries. Due to this, the characterization of TZV will not be tackled in detail in this work. However, it will be exploited in the last part of this Thesis, since TZV is fundamental in SRAM PUFs, as will be explained in following Sections.

1.3 Time-Dependent Variability

Since the early years of microelectronics, MOSFET technology has been advanced by a continuous dimension scaling [2]. Such scaling has led to numerous improvements in aspects such as transistor density, switching speed or power dissipation.

However, this dimension scaling in CMOS technologies also brings new concerns in terms of TDV phenomena. For instance, this size shrink does not come with a constant field scaling as the transistor dimensions enter the sub-micron region [20]. This leads to higher electric fields within transistors, which can degrade the transistor parameters, and thus the circuit performances, during the circuit operation. This complication induced by technology scaling is not only quantitative, but also qualitative. As it will be explained in the following, several of the most important TDV phenomena, such as Bias Temperature Instability, Hot Carrier Injection and Random Telegraph Noise, involve the trapping and detrapping of charges in individual defects present in the transistors. When the transistor dimensions are large enough (e.g., in the micrometer range), a very large number of such defects is present in each transistor. Therefore, although the exact number may not be the same, these phenomena can be modeled as deterministic and continuous. However, in small-area devices (e.g., in the nanometer range), only a handful of defects may contribute to these phenomena in each transistor [21]. Therefore, different devices may display a very different behaviour, and these phenomena must be modeled as stochastic and discrete. This leads to a paradigm shift in terms of reliability characterization, modeling and simulation, since identical circuits, subject to exactly the same operation conditions, may degrade in different manners.

TDV phenomena can be divided into transient effects, such as RTN, and degradation phenomena, such as BTI, HCI and Time-Dependent Dielectric Breakdown (TDDB). While the last one is out of the scope of this thesis, a more detailed explanation of RTN, BTI and HCI is presented below.

1.3.1 Random Telegraph Noise

The Random Telegraph Noise phenomenon is observed as a random switching of the transistor drain current between two or more discrete levels along time. These fluctuations are caused by the trapping/detrapping of charge carriers in/from de-

fects present in device oxide and interface (see Fig. 1.2), which create shifts in the transistor parameters, such as V_{th}. The defects that cause RTN can be present in the transistor right after it has been fabricated, or can be generated afterwards, during device operation, by degradation mechanisms [22].

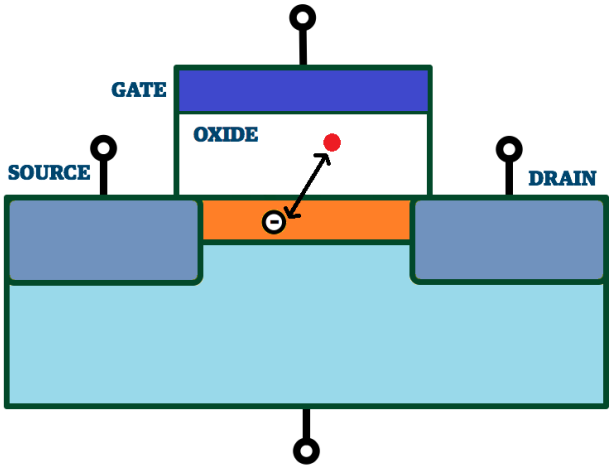

Figure 1.2: Representation of the trapping/detrapping of charge carriers that originate RTN.

The impact of RTN on circuits as a source of performance variability has been widely studied, for example in SRAMs and Ring Oscillators [23], image sensors [24], or flash memories [25]. Furthermore, this impact is increasing as the devices continue to be scaled down [26]. It has also been shown that the negative effect of RTN on circuits increases as the supply voltage decreases, which can make it threatening in low power applications [27].

Fig. 1.3 displays two examples of current traces that display RTN behavior. It can be seen how, depending on the number of involved defects that trap/detrap charge carriers, the number of discrete current levels vary.

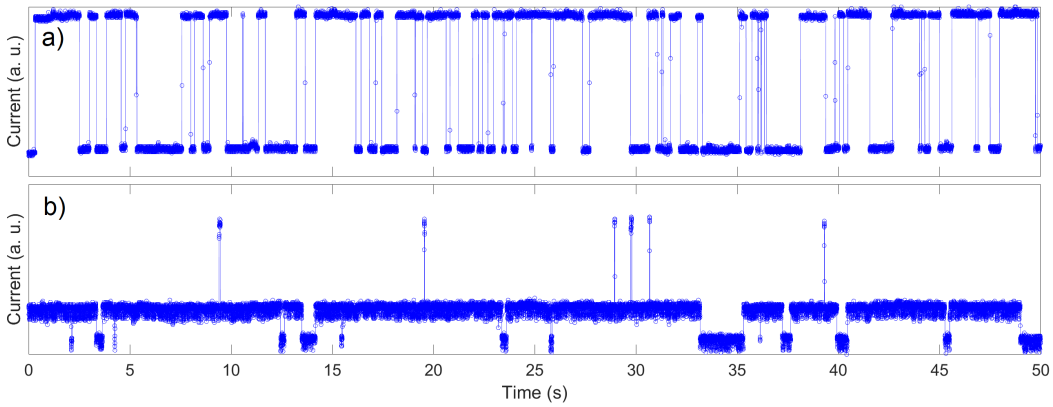

Figure 1.3: Two current traces displaying RTN. In a), only one defect is giving rise to detectable RTN; in b), two.

In particular, the number of detectable RTN defects can be inferred from the number of discrete current levels as [28]:

$$N = ceil(log_2 N_L) \tag{1.1}$$

where N is the number of detectable RTN defects, and N_L is the number of current levels in the trace.

In fact, the RTN can vary largely from one device to another, especially in the case of deeply-scaled technologies, in which each transistor contains a more or less reduced number of defects, which is random and not equal for all devices. In addition to the number of defects, the parameters that characterize RTN are the current amplitude of the fluctuations associated to each defect (ΔI), or analogously the threshold voltage shifts (ΔV_{th}), and its associated time constants. Notice that both the associated shifts and the time constants of each given defect are also stochastic variables that can be modeled with some statistical distributions. For example, the distribution followed by the threshold voltage shifts associated to the trapping/detrapping of defects has been described as exponential [29] or lognormal [30] in the literature. The time constants are the capture time (τ_c), which is the average time that an RTN defect takes to capture a charge carrier when it is empty ($<t_c>$), and the emission time (τ_e), which is the average time that a defect takes to emit the charge carrier once it is occupied ($<t_e>$) [1]. It is important to remark that these variables represent indeed only the average values, and that the actual time-to-capture (t_c) and time-to-emission (t_e) for each trapping and detrapping event of a given defect are stochastic values that follow a distribution described by [31]:

$$p(t_e) = \frac{1}{\tau_e} \cdot e^{\frac{-t_e}{\tau_e}} \qquad p(t_c) = \frac{1}{\tau_c} \cdot e^{\frac{-t_c}{\tau_c}} \tag{1.2}$$

This means that the actual t_e and t_c times for each emission/capture event of a defect with a given τ_e and a given τ_c can span across several orders of magnitude in time. Additionally, the time constants that characterize a given defect are dependent on the operation conditions, namely on the bias voltages of the transistor [32], and on the temperature.

Regarding the biasing conditions, the capture (emission) time constant is expected to decrease (increase) as the absolute value of the gate voltage is increased, following an exponential law [32], [33]:

$$\tau_e = \tau_{e\_0} \cdot 10^{\beta_e V_{gs}} \qquad \tau_c = \tau_{c\_0} \cdot 10^{\beta_c V_{gs}} \tag{1.3}$$

where τ_{e_0} and τ_{c_0} are the values that the emission and capture time constants of a given defect take when there is no voltage difference between the gate and source terminals of the device.

[1] The "$< >$" symbols will be used throughout this Thesis to indicate a mean value.

This means that, at higher absolute values of the gate voltage, defects have a higher tendency to be occupied by a charge carrier. On the other hand, when temperature is considered, the trapping and detrapping behavior of traps can be modeled as an Arrhenius process [34], which implies that both the capture and the emission time constants of a given defect are expected to decrease with temperature according to

$$\tau_e = \tau_{e\_0} \cdot e^{\frac{E_{a\_e}}{k_B T}} \qquad \tau_c = \tau_{c\_0} \cdot e^{\frac{E_{a\_c}}{k_B T}} \tag{1.4}$$

where $E_{a\_e,c}$ is the activation energy of the emission/capture process, k_B is the Boltzmann constant, and T is the temperature. Therefore, both the trapping and detrapping behavior is expected to be accelerated through an increase in temperature.

1.3.2 Bias Temperature Instability

Bias Temperature Instability is a gate-voltage and temperature activated TDV phenomenon that causes a progressive degradation over time of the transistor parameters, e.g., an increase in the absolute value of its threshold voltage. This can become a critical issue for circuit functionality, especially for technologies in the nanometer range [4].

Negative Bias Temperature Instability (NBTI) in PMOS transistors, which takes its name from the negative bias at the transistor's gate that causes it, has been traditionally the main concern. However, Positive Bias Temperature Instability (PBTI), caused by positive bias in the transistor's gate in NMOS transistors, has gained importance since the introduction of High-k Metal Gate dielectrics (HKMG) [35], [36].

Since the first work proposing a hydrogen-diffusion controlled interface state creation mechanism in 1977 [37], most of the published works seemed to support such a reaction-diffusion (RD) theory. However, in 2006, it was suggested that BTI could also be caused by the trapping of charge carriers into defects present in the device oxide [38]. Since then, several studies support that charge carrier trapping is the main contribution to BTI degradation [39], [40]. Furthermore, nowadays there is a general agreement around the fact that many of the defects responsible of BTI can also generate RTN [41]. In this sense, Fig. 1.4 depicts a schematic representation of the charge carrier trapping phenomenon that causes BTI degradation. In older technology nodes, the number of defects in each transistor was large enough to consider the BTI phenomenon as deterministic; identical devices would degrade the same under the same workload condition. However, in deeply-scaled technologies, the decrease in the number of defects lead to a more stochastic behavior, which means that different devices designed identically may degrade differently under the same workload condition. This occurs because different transistors may have a different number of defects and, also, the parameters associated to each defect can take different values. As in the case of RTN, these parameters are the time constants associated to each defect, and the associated current or threshold voltage shift.

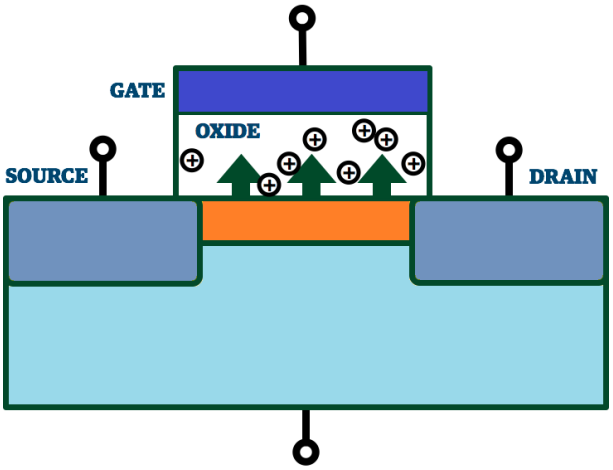

Figure 1.4: Representation of the trapping of charge carriers that originate BTI.

The degradation caused by BTI can be divided into a permanent or quasi-permanent component and a recoverable one [42], [43], [44]. While the recoverable component recuperates gradually in the time scale of the experiment once the large gate voltage decreases, the (quasi-) permanent one stays, or at least it takes a much longer time to recuperate. For transistors of "large" dimensions (e.g., in the micrometer range), both the initial degradation and the following recovery are seen as "deterministic" [2] and continuous phenomena. However, in smaller devices (i.e., in the nanometer range), both the initial degradation and the following recovery are seen as discrete phenomena, since each defect that captures/emits a charge carrier will produce a discrete shift in the transistor parameters.

1.3.3 Hot Carrier Injection

Hot Carrier Injection has been studied in the last decades as a source of circuit degradation [45], [46]. In the last years, this phenomenon is becoming of growing interest due to the increasing electric fields that the continued scaling of the devices has brought [47].

The HCI phenomenon occurs when the transistor is on (V_{GS} is larger than the threshold voltage) and there is a lateral field created by V_{DS}, which reaches its maximum in the region near the drain. This lateral field causes the acceleration of the charge carriers and the injection of these high-energy carriers (hot carriers) into the gate dielectric. Unlike the case of BTI, in which the vertical field leads to a uniform damage across the gate-oxide area, the carrier acceleration caused by the lateral field in the case of HCI leads to a non-uniform damage across the channel. In particular, the largest part of the damage caused by HCI is located close to the

[2]Notice the utilization of quotation marks. TDV phenomena are not actually deterministic for "large" technologies, since the exact number of defects may still be different from device to device, and the associated amplitudes and time constants of those defects also follow statistical distributions. However, due to the high number of defects, the differences are "averaged out".

drain area [48]. Fig. 1.5 depicts an schematic representation of this injection of hot carriers into the dielectric.

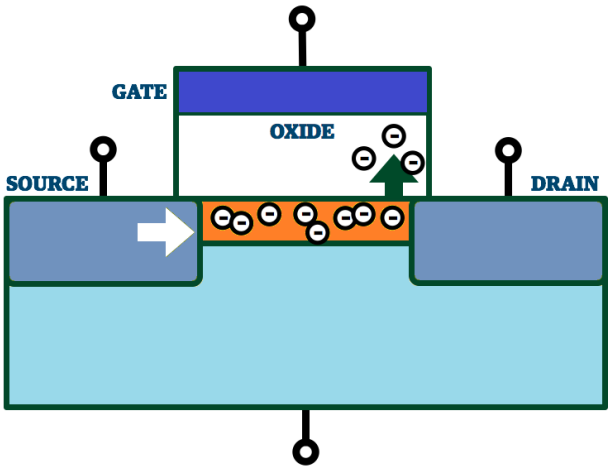

Figure 1.5: Representation of the injection of hot carriers into the oxide that originates HCI.

Analogously to the case of BTI, the degradation caused by HCI reveals a stochastic behavior for technologies in the nanometer range. This degradation causes the quasi-permanent deterioration of the transistor parameters, such as the threshold voltage or the mobility.

1.4 A model for Time-Dependent Variability: the Probabilistic Defect Occupancy model

During the last decades, very different approaches have been followed to model TDV. For example, authors in [49] describe the degradation of the device threshold voltage through a number of equations based on physical parameters. This analytical approach to the evolution of the degradation leads to a deterministic prediction of the device degradation (i.e., devices subject to the same operation conditions are predicted to have the same degradation). Although this may be a very good approximation when older technology nodes (i.e., larger transistor sizes) are considered, that is not the case for deeply-scaled technologies in the nanometer range, in which a clear stochastic behavior is observed for the degradation [50].

Another approach relies on the utilization of Technology Computer Aided Design (TCAD) models [51], [52], [53]. These models focus heavily on describing the device physics, such as the current flow and the heat generation in a transistor. Although this may be very useful to obtain a deep insight into the physics involved in the different TDV phenomena, the complexity associated to these models advise against their utilization in circuit simulation.

Another approach to model BTI is the Probabilistic Defect Occupancy model, which

was introduced in [10] and is based on the trapping/detrapping of charge carriers in/from defects that exist in transistors. Because of this, the PDO model is suitable to emulate the discrete and stochastic behavior that TDV present in nanometer-range technologies. Furthermore, the PDO model is relatively simple, especially when compared to more complex tools such as TCAD models, and makes use of only a handful of parameters, which allows its utilization in computationally efficient simulators.

First, the case of a transistor with a single defect can be considered. The basis of the PDO model is that the trapping/detrapping events undergone by a defect can be modeled through a memoryless random process called Markov process [54], i.e., it is a random process that only depends on the present situation at any given instant, and not on the past conditions. In the particular case of trapping/detrapping events, this means that, at every instant Δt, the defect has a probability to capture a carrier $P_c = \Delta t/\tau_c$ if it is empty, and a probability to emit it $P_e = \Delta t/\tau_e$ if it is occupied, as long as τ_c, $\tau_e \gg \Delta t$. In those formulas, τ_c and τ_e represent the capture and emission time constants, as defined in Subsection 1.3.1. These parameters are bias and temperature dependent. The other parameter that characterizes the defect is η, which represents the threshold voltage shift that the trapping/detrapping event produces. Then, in this simple scenario, in which only one defect is present in the device, it would be enough to characterize its parameters τ_c, τ_e and η, together with their voltage and temperature dependences, to simulate the evolution of the device under various operation conditions.

In a real device, however, more than one defect may be present. To account for this, the total shift in threshold voltage for a device with N defects with respect to its threshold voltage when no defect is occupied, can be formulated as

$$|\Delta V_{th}(t)| = \sum_{j=1}^{N} k_j(t) \cdot \eta_j \tag{1.5}$$

where j is an index that denotes each defect in the device, and $k_j(t)$ can be 1 or 0 if the j-th defect is occupied or empty at time t.

Therefore, three different types of parameter distributions must be extracted from the characterization experiments to construct the PDO model:

- The distribution of the threshold voltage shift amplitudes η associated to the trapping/detrapping of defects.
- The distribution of the emission and capture time constants (τ_e, τ_c), which determine the probability of occupation $P_{occ,j}$ of each defect under certain operation conditions at a given time instant, and therefore its k_j.
- The distribution of the number of defects. There are two possible approaches to this. First, it is possible to use the total number of defects present in the device. Then, by assigning time constants (τ_e, τ_c) to each defect, it is possible to predict if they will be "active" (i.e., trap/detrap charge carriers) during

a given time window at certain conditions. A more empirical approach is to directly characterize the number of active defects at given conditions.

Additionally, the dependence of some of these parameters, such as the defect time constants and the threshold voltage shift amplitudes, on the operation conditions must be taken into account.

Once all these parameters have been successfully characterized, the PDO model can be fully built and integrated in a stochastic reliability simulator such as the one presented in [55].

1.5 A complete flow towards reliability-aware circuit design

Once the main TZV and TDV variability phenomena in CMOS technologies have been described, it becomes obvious that their impact on circuits must be accounted for during the design process. However, this is not a direct task, and instead must be performed in a number of steps, which are depicted in Fig. 1.6.

First, variability phenomena must be characterized. Both TZV and TDV display a stochastic nature in deeply scaled technologies. Because of this, their characterization must be done in a massive manner (i.e., hundreds or thousands of devices must be measured) to obtain statistically relevant information. Furthermore, this characterization must be performed at different conditions in order to study the different variability phenomena, which leads to the generation of enormous amounts of data. Therefore, the manual analysis of such amounts of data becomes unfeasible, and the development of automated analysis tools becomes fundamental. As it has been already mentioned, nowadays the semiconductor foundries provide models for the TZV of the technologies that they fabricate. Therefore, the focus of this Thesis is rather set on the statistical characterization of TDV phenomena.

Once that the characterization process has been completed, the parameters extracted during that phase can be used to construct models that describe the variability phenomena. Again, the focus here will be set on TDV models, since TZV models are provided by the foundries. In this Thesis, the Probabilistic Defect Occupancy (PDO) model is used [10]. This model follows a stochastic and defect-centric approach of TDV and assigns discrete shifts of the transistor parameters to the trapping/detrapping of charge carriers into defects.

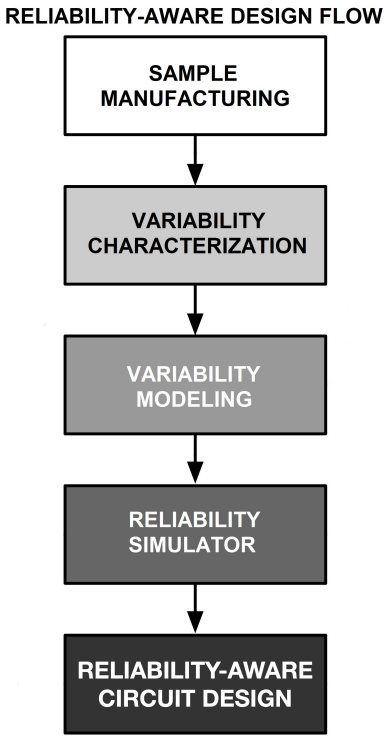

Figure 1.6: Block diagram for the development of RAD solutions.

Once that such a stochastic model for TDV has been constructed, it can be integrated, together with a model for TZV, into a stochastic reliability simulator. Such a simulator should predict in an accurate manner the variability expected for a circuit under certain operation conditions. One such a circuit-level simulation methodology for RTN has been presented in [56]. In it, the effect of charge carrier trapping/detrapping from individual defects in the transistors is mimicked through the inclusion of a variable DC voltage source at the gate of each device. Then, if the RTN parameter distributions have been correctly characterized and the corresponding model has been constructed, the impact of RTN on circuits will be correctly simulated.

Tools that aim at the simulation of the degradation phenomena have also been proposed. For example, [57] presents a deterministic aging simulator for BTI and HCI. However, this deterministic approach can prove inadequate when dealing with deeply scaled technologies, in which these phenomena reveal a stochastic nature. CASE, a simulation tool that accounts for that stochastic nature by using the PDO model, has been presented in [55]. Some of the most important features of this tool, apart from its ability to handle the stochasticity of the variability phenomena, are the possibility of including the effect of TZV and TDV in a combined manner, the inclusion of the bi-directional link between stress conditions and degradation, or the possibility to perform the simulations using a size-adaptive time-step to account for such link, which can highly increase the simulation efficiency.

Then, once that accurate simulation tools for variability phenomena have been developed, they can be used by circuit designers to study and take into account the impact that those phenomena have on a given design. These can be done in different

ways. For example, in [58], an optimization methodology that aims at the generation of lifetime-aware Pareto-optimal fronts has been presented. This methodology is based on the inclusion of the CASE simulator [55] in a multi-objective optimization flow. In it, the circuit lifetime (closely related to the concept of the reliability) can be optimized together with other circuit performances, which allows the designer to achieve the best performance trade-offs for the longest possible lifetime. Another example of Reliablity-Aware Design that follows a very different approach has been presented in [59]. In it, an on-chip dynamic reliability management technique to monitor BTI degradation in an SRAM, and compensate it, has been presented.

1.6 Exploiting variability in hardware security applications

1.6.1 Why PUFs?

Up to this point, some of the main variability phenomena in CMOS technologies have been described. In particular, the focus has been set on Time-Dependent Variability phenomena, since semiconductor foundries usually provide accurate models to account for Time-Zero Variability. This has been done by following a clear approach: to characterize and model these TDV phenomena so that they can be accurately simulated, and therefore RAD solutions can be developed to mitigate their negative impact on circuits. In this Section, the focus is rather set on TZV, and the approach is completely reversed: variability in CMOS technologies is exploited for the user's benefit through the utilization of PUFs. However, as it will be explained later, it must be remarked that the exploitation of TZV in hardware security applications does not imply that the degradation of these circuits due to TDV phenomena can be neglected.

The emergence of the concept of PUFs is closely related to the growth of the Internet of Things (IoT). The basic idea behind the IoT is the existence of a massive number of physical objects such as Radio-Frequency IDentification (RFID) tags, sensors, actuators and cell phones, which, through unique addressing schemes, are able to interact with each other and cooperate with their neighbors to reach common goals [60]. These devices can be deployed in very diverse fields, such as in the medical and healthcare area[61], in the implementation of smart homes [62], in industry [63], or in automotive [64], among others.

Although IoT can bring improvements in all those diverse aspects of our daily lives, it also gives rise to some challenges. For instance, enormous amounts of data are gathered and transmitted by the IoT sensor nodes. These data can contain sensitive, or even safety-critical information, and therefore the communication channels between all the different IoT devices must be secured.

In this context, PUFs act as a root of trust of those IoT devices. Frequently, they make use of inherent TZV in CMOS technologies to generate unpredictable bits

that can be used in different security applications, such as anti-counterfeiting, device authentication or key generation.

1.6.2 What are PUFs?

The basic idea behind PUFs was introduced in [65] under the name of Physical One-Way Functions. In general, a PUF can be defined as a device that exploits inherent randomness introduced during manufacturing to give a physical entity a unique 'fingerprint' or trust anchor [66]. That randomness is inherent to the manufacturing process and cannot be avoided. Furthermore, it is not possible to model, predict or replicate it, even by the manufacturer. It must be noted that there have been some PUF designs that are not based in manufacturing-induced randomness, such as the soft oxide breakdown PUF presented in [67]. However, these represent only a small minority of the overall existing PUF designs.

As its name indicates, Physical Unclonable Functions perform a functional operation: it receives an input (usually denoted challenge) and produces an output (response). Each input together with its corresponding response receives the name of challenge-response pair (CRP). Due to the impossibility to predict, model or replicate the manufacturing-induced randomness, and to the fact that the response to any given challenge is based on this randomness, the response to any given challenge is unique to each fabricated PUF and remains hidden from any potential attacker.

There are several properties that candidates should fulfill to be considered a valid PUF [68]. The most important ones are:

- **Unclonability:** it should not be possible to clone a PUF, either physically or mathematically. This should be true even for the original manufacturer of the PUF. This is achieved through the basic operation principle of the PUF, which relies on the random variations that occur during the manufacturing process.
- **Unpredictability:** the response of a PUF instance to a given challenge should be impossible to predict, even if the user has information about part of the response, or about the response of a different PUF instance to that same challenge.
- **Uniqueness:** this is perhaps one of the most basic properties of PUFs. Since PUFs act as the circuit's fingerprint, it is utterly important that different PUF instances have different responses to the same challenge. Fig. 1.7 shows a schematic representation of this property.

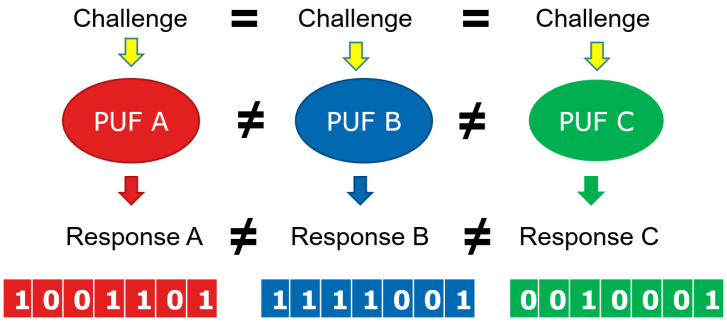

Figure 1.7: Representation of the property of uniqueness of PUFs.

- **Reliability:** since the response of a PUF instance to a given challenge may be read multiple times, this response should always stay constant. This can be challenging specially when factors such as circuit aging, supply voltage variations or temperature variations are considered. Fig. 1.8 portrays a schematic depiction of this property.

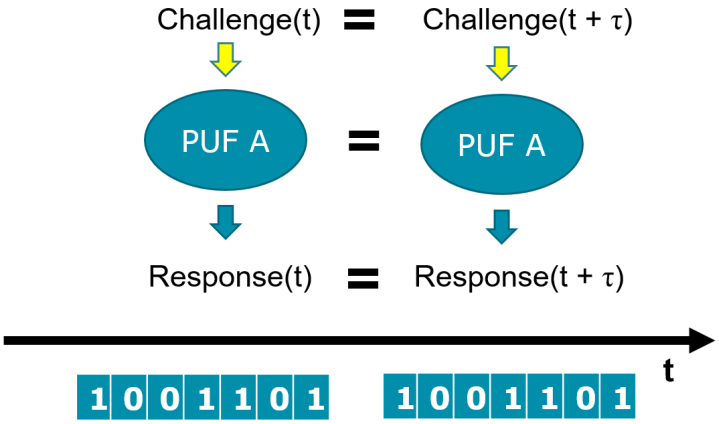

Figure 1.8: Representation of the property of reliability of PUFs.

1.6.3 PUF applications

PUFs have a wide variety of applications within the field of security. In the following, some of the most important ones are discussed.

1.6.3.1 PUFs for entity authentication

As it has been already mentioned, PUFs exploit unpredictable variations occurred during the manufacturing process to identify specific devices [69]. PUF-based device authentication consists of two phases: the enrollment and the authentication phase. These can be seen in Fig. 1.9.

During the enrollment phase, which occurs only once and must be performed in a secure environment, a challenge is sent to the PUF, and the corresponding response is then received. The CRP is then recorded as a reference for authentication.

Later, for a device that claims to be the registered one, the authentication phase is performed. Logically, this phase can occur several times. During the authentication phase, the challenge from the database is sent to the device that needs authentication. Then, the response returned from that device is compared to that from the CRP recorded during the enrollment phase. After the CRP has been evaluated, the server can validate the authentication of the device if the response is correct or reject it if it is erroneous. Until here, only 1 CRP has been considered for each PUF. This has been done for the sake of simplicity. However, this is not always the case. In fact, PUF designs can be divided in two distinct categories according to the number of possible CRPs that they can provide. In this sense, weak PUFs are characterized by having only one or a small number of challenges, while strong PUFs can support a large enough number of challenges such that the complete determination of all CRPs within a limited time window is not feasible [70]. Finally, it must be noted that entity authentication can be used to validate the identity of a device, or to avoid chip counterfeiting [71].

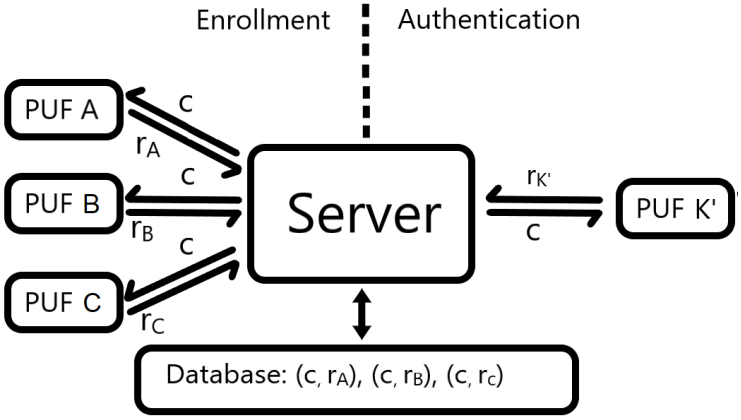

Figure 1.9: Block representation of the enrollment and authentication phases of a PUF used for device authentication.

.

1.6.3.2 PUFs for key generation

Secret keys are essential in many cryptographic applications to perform encryption and decryption. The key is generated through a True Random Number Generator (TRNG), and a consecutive key generation function, which ensures the quality of the generated key. This key should be secret for the system to be secure, according to the Kerckhoffs's principle [72]. Traditionally, the secret key is stored in a Non-Volatile Memory (NVM) such as a flash memory, an e-fuse or an anti-fuse, among others. The content of the NVM should not be accessible by an attacker by any means. Unfortunately, this is not always the case, since it is possible to apply different attack techniques to the NVM [73], [74]. Additionally, the key must be generated by a secure party.

In this context, PUFs have been broadly discussed to improve the generation and storage of secret keys. On the one hand, in terms of the key generation, PUFs offer a more secure option because the entropy is not supplied by an external source. On the other hand, in terms of key storage, PUFs offer another improvement, since the keys are not visible when the device is powered-off, and therefore they present a higher resilience against physical attacks [75]. Weak PUFs that only provide a single CRP are usually preferred for key generation, since only one key is required in cryptographic algorithms [69].

1.6.3.3 PUFs for True Random Number Generation

Another important application of PUFs is their utilization as TRNG [76], [77]. True Random Numbers are a collection of bits that are unpredictable and display statistical properties of randomness and are a fundamental part of many cryptographic applications. If such numbers can be guessed with any accuracy, the security of the cryptographic application will fail [78].

The main requirement for a PUF to successfully work as a TRNG is the complete opposite of the one needed for entity authentication or for key generation: the PUF should return a completely random response each time that a given challenge is applied. This may seem contradictory with some of the other PUF properties. However, as will be seen in later Sections, this conflict can be solved in some types of PUF implementations that have two different types of bit cells (i.e., cells that provide one bit of the PUF response): those that ideally always provide the same bit response to a given challenge, which can be used for entity authentication or key generation, and those that sometimes provide one bit value and other times the opposite value in a random manner, which can be used for true random number generation. This will be explored in further detail in Section 1.7 and, later, throughout Chapter 5. The adequacy of the generated response in terms of randomness and unpredictability can be assessed by the tests created by the National Institute of Standards and Technology (NIST) [79].

1.6.4 Non-ideal behavior of PUF reliability

Up to this point, PUFs have been considered ideal in terms of the reliability of their response. That is, according to the schematic representation of the reliability property depicted in 1.8, a given PUF instance will always return the same response when a given challenge is applied for entity authentication or key generation. However, in reality, this is not always the case. In the case of silicon-based PUFs (that will be introduced in the following Section), there are a series of factors that can induce errors in the PUF response. The main ones are noise, bias and temperature variations, and TDV phenomena that can degrade the PUF circuit. These factors can cause some of the bits from the PUF response to be different to the ones stored in the server database, which would result into a failed entity authentication or key generation. Because of this problem, the enrollment and authentication phases are not as simple as shown in Fig. 1.9. Instead, PUFs usually incorporate Helper Data

Algorithms (HDA) to solve this reliability problem. A typical HDA is an assembly of components rather than a unitary block [80]. In particular, there are usually three different steps involved in HDA. The first phase is bit selection. In it, the least reliable bits are discarded. This first step eases the second one, which is the correction of errors. After this, entropy compression is performed to create a key with the optimum entropy. Although there are different implementations of the HDA [81], [82], the most important takeaway is that the complexity of the associated circuitry is proportional to the ratio of errors that must be corrected [83], which highlights the importance of reducing such errors as much as possible prior to the error correction phase.

1.6.5 PUF implementations

The first PUF designs proposed in the literature were not electronic PUFs and followed very diverse approaches [69]. Some of them are the Optical PUF [65], which makes use of the unique interaction of a laser beam with a scattering medium; the Paper PUF [84], which uses the unique fiber structure of paper documents, or the CD PUF [85], which exploits the inherent variability of the lengths of the lands and pits of compact disks as the PUF fingerprint.

However, electronic or silicon-based PUFs are in general easier to implement, and they have gained importance in the last years. Very diverse such silicon-based PUF implementations have been presented. One of the first proposed techniques for unique identification using this type of PUF, without the need of any special processing steps, was the V_{th}-PUF, proposed in [86]. This PUF includes a number of identical transistors, each of which driving a resistive load. Due to the impact of TZV on the transistors' characteristics, and in particular on the transistors' V_{th}, the current flowing through each load is different, fact that was exploited by the PUF. The authors of [87] introduced the Arbiter PUF. The basic idea for the CRP of this PUF is to perform a digital race between two ideally identical paths in the chip, and to evaluate which of the two paths is faster with an arbiter circuit. In this way, the impact of TZV in the delay of the paths is exploited. Another implementation is the Ring Oscillator PUF, which was first presented in [88], [89]. The basic idea in this case is to compare the oscillation frequencies between a pair of ideally identical ring oscillators. These frequencies will be different due to TZV. Another family of electronic PUFs is formed by the Memory-based PUFs [76], [90], [91]. Among them, the most important one is the SRAM PUF. This type of PUF is the one studied in the last part of this Thesis and is described in greater detail in the next Section.

1.7 The SRAM PUF

The Static Random Access Memory (SRAM) PUF was first presented in [76]. SRAM-based PUFs can result very convenient since SRAM memories are ubiquitous circuits, and it is possible to use a fraction of the cells that compose a general-purpose SRAM for the PUF implementation, instead of devoting purposely-designed

circuitry. The idea behind this type of PUF is to use the power-up state of Static Random-Access Memory (SRAM) cells as an identifying fingerprint of circuit instances. The core of a conventional 6T SRAM cell, shown in Fig. 1.10, is formed by two cross-coupled inverters. Therefore, there are two possible stable states that can be stored by the cell, labelled as '0' and '1'. Throughout Chapter 5 of this Thesis, the '0' state will correspond to the left node Vl being at a low voltage and the right node Vr at a high one, and the '1' state will denote the inverse situation. The cross-coupled inverters constantly reinforce the stored value, which, ideally, is held indefinitely as long as power supply voltage is applied. Furthermore, the stored value can be accessed, either to read it or to write a new one, by activating the access transistors ($Mnal$ and $Mnar$ in Fig. 1.10). Although the utilization of SRAM cells in the context of PUFs has some differences with respect to its conventional utilization as memory cells, the main characteristics of this conventional operation will be outlined first, since they can be very useful to understand the basic working principles of the cell.

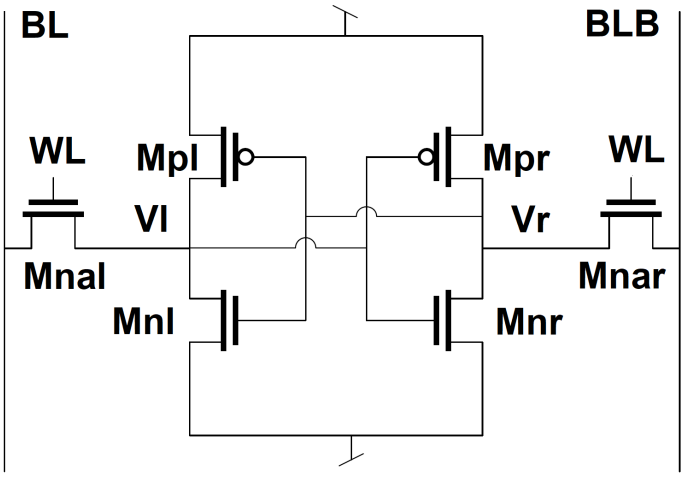

Figure 1.10: Schematic of a 6T SRAM cell.

.

To write a new value on the cell, the bitlines BL and BLB, corresponding to the nodes Vl and Vr respectively, are set to the desired logic level, and WL is activated. For the sake of illustration, consider that the SRAM cell has a '1' stored (voltage Vl is high and voltage Vr is low). Then, if the opposite value is to be written, BL is set to logic zero and BLB to logic one. When WL is activated, the access transistors are switched on, the inverter is pulled past its trip point and causes the cell to latch and hold the new value. To read the value stored in the cell, both bitlines are driven to a logic one and are kept floating. When WL is activated, the bitline close to the cell node storing a logic zero starts to discharge. Then, by detecting which of the bitlines has discharged, it is possible to assess which is the value written on the cell.

For a good read stability, the pull-down devices (Mnl and Mnr in Fig. 1.10), must be stronger than the access transistors ($Mnal$ and $Mnar$ in Fig. 1.10). On the other hand, the writeability increases when the access transistors are stronger than the pull-up transistors (Mpl and Mpr in Fig. 1.10). These considerations, together

with the minimization of device sizes for maximum packing density yields some usual sizing criteria: minimum widths and lengths for pull-up and access transistors and twice this ratio for the pull-down devices.

Until here, the basic working principles of an SRAM cell have been described, namely the writing and reading operations. But how do PUFs exploit the TZV of SRAM cells to generate a unique fingerprint of the circuit? Let us consider the situation in which a given cell is powered up and the access transistors remain turned off. In principle, since the access transistors remain turned off, no value is externally written on the cell. Therefore, one could imagine that the cell would remain in a metastable intermediate state between '0' and '1', since the core inverters are identical by design. However, in reality, the cell will evolve towards one of the two states in an unpredictable manner. This occurs because, although the cross-coupled inverters are ideally identical, unavoidable TZV occurs during the manufacturing phase. Although the four core transistors of the cell play a role in determining the preferred power-up state, let us consider first, for the sake of simplicity, an SRAM cell in which both pull-down NMOS transistors are identical, and only the asymmetry between the PMOS devices come into play during the power-up. If the left PMOS device Mpl is stronger (i.e., has a smaller $|\mathrm{V_{th}}|$) than the right one, it will conduct faster during the power-up, and will therefore pull the left node Vl faster. Therefore, the cell will have a preference to power up towards the '1' state. If the four core transistors are considered, an analogous reasoning can be followed, in which the PMOS pull-up transistors compete during the power-up to pull their corresponding nodes up, while the NMOS pull-down transistors compete to pull their corresponding nodes down. In fact, the Mismatch Factor metric [92], constructed from comparing the threshold voltages of the core transistors as

$$MF = (|V_{th\_pl}|-|V_{th\_pr}|) - (V_{th\_nr} - V_{th\_nl}) \tag{1.6}$$

is strongly correlated to the preferred power-up values of the cells. Therefore, SRAM cells that have a considerable asymmetry between their inverters have a stronger preference to power up to one of the two possible states. This property is exploited for device authentication or secret key generation by PUFs based on power-up SRAM states. This idea has not only been studied in academia, but it has also been adopted by the industry [75]. On the other hand, the cells that do not have a large asymmetry between their inverters, and therefore do not have a well-defined preferred power-up state, can be used for TRNG, since they will sometimes power up to '0', and sometimes to '1', in a random and unpredictable manner.

In general, a PUF instance used for entity authentication or secret key generation should always return the same response for the same challenge. This is naturally also true for SRAM-based PUFs. However, as it has been explained in Section 1.6.4, this is not always the case: factors such as temperature, noise (like RTN) or circuit aging may undermine the reliability of a PUF instance. Although the utilization of error-correction techniques can correct, to a certain extent, erroneous PUF responses [70], the overhead associated to these components grows sharply with increasing error-correcting capability [93]. Therefore, it is important to maintain the error rate as low as possible by enhancing the reliability of the PUF instance itself.

One of the approaches to increase the reliability of SRAM PUF responses is the Dark-Bit Masking [94]. The idea behind Dark-Bit Masking is to mark as "dark" the PUF bits that may display an unstable (i.e., not reliable) response. In the case of the SRAM PUF, this would mean to mark as dark the SRAM cells that do not always display the same power-up state. Then, these dark bits are not included in the PUF readouts, so that they do not contribute with errors to the PUF response. In this way, the reliability of the PUF is highly increased.

In summary, methods for improving the reliability of SRAM PUF responses are critically required.

1.8 Main contributions of this Thesis

This Section enumerates the main original contributions of this Thesis. For a more detailed explanation of the content of each Chapter, refer to Section 1.9. Those main contributions are:

- TiDeVa, a toolbox for the automated analysis and parameter extraction from TDV tests is presented. This toolbox is based on the Maximum Likelihood Estimation method and allows the extraction from massive amounts of experimental data without any user supervision.
- The different parameters that characterize RTN defects are extracted with TiDeVa at different operation conditions. Special interest is set on the determination of the distribution of the current shift amplitudes associated to charge carrier trapping/detrapping in/from defects, as the other defect parameter distributions, such as that of the time constants, will be tackled through the aging tests.
- A new RTN analysis methodology, based on the Maximum Current Fluctuation metric, is presented. This method overcomes some of the drawbacks of more conventional approaches. In particular, it does not require of any complex analysis, and it is able for account for defects with very small amplitudes that may not be detected by those conventional approaches.
- The impact that the biasing conditions that devices experience prior to their measurement is studied. It is seen how this factor can largely impact the measurement results, and thus should not be overlooked.
- Aging tests are performed and analyzed with TiDeVa, and the necessary parameters to tackle the construction of a TDV model are extracted by the tool.
- Using the paramters extracted from the aging tests by TiDeVa, a complete methodology to extract the distribution of the time constants of the PDO model and the number of defects per device is thoroughly described, and the corresponding results are presented.

- The extracted information about the number of defects, their associated amplitude distribution, and their time constant distribution are used to construct the PDO model, which can be then integrated into an RTN simulator. This simulator allows to predict the impact of RTN at the device and at the circuit level.
- Again, using all that information, an emulator of the RTN experimental data is constructed. This tool allows to generate RTN traces as they would be recorded in the laboratory, which can be very useful in order to test new characterization strategies in a much faster and cheaper manner than if actual experimentally measured traces were used.
- A chip, the KipT chip, that contains an SRAM cell array, is presented. This chip does not only allow the conventional write/hold/read operation of SRAMs, but also allows the custom application of accurate voltages at the different terminals of the cells, which can be very useful, for example, to apply controlled stress to the circuit.
- A method is presented to improve the reliability of SRAM PUFs. This method, the MTSV method, aims at selecting the cells with a stronger power-up tendency not only at nominal operation conditions, but also when factors such as temperature or voltage variations, as well as circuit degradation, are considered.

1.9 Structure of the rest of this Thesis

Chapter 2 of this Thesis presents a complete TDV characterization framework. This includes a detailed description of the characterization strategy for both RTN and aging phenomena, and a explanation of the information that needs to be extracted from each type of test. Then, according to the presented characterization strategy, a number of requirements that the experimental setup used in this task should fulfill are enumerated. These requirements are addressed in the following Sections, which are devoted to a description of the Endurance chip [95], a transistor-array chip that has been used for the characterization of TDV at the device level, and to a custom experimental setup that allows, together with the Endurance chip, the accurate and automated performance of those TDV tests. Finally, this Chapter is concluded with the presentation of TiDeVa, a software tool that allows the automated analysis and parameter extraction from the data generated during the TDV tests. This tool, which makes uses of the Maximum Likelihood Estimation method, has a simple and user-friendly Graphical User Interface that facilitates the analysis process. Some examples of the RTN and aging phenomena characterization performed by TiDeVa are provided at the end of the Section.

Chapter 3 focuses on the characterization results of RTN. In particular, the most important information extracted from the RTN tests consists in the distribution of the amplitudes associated to the RTN defects trapping and detrapping. These will be used in the construction of the PDO model. After that, a new method to characterize the RTN phenomenon, based on the Maximum Current Fluctuation metric,

is presented. This method could be an interesting alternative to more conventional analysis techniques, as it overcomes some of their drawbacks, such as requiring very complex processing techniques, or not being able to detect defects with a very small associated amplitude. Finally, Chapter 3 is concluded with the analysis of a problem that is often overlooked: the impact that the biasing conditions that the devices experience prior to the actual measurements can have on the outcome of the measurement.

Chapter 4 tackles the problem of determining the distribution of the time constants of defects in the PDO model. This task, that proves to be not at all straightforward, is first approached through a path that, although not very efficient, will prove useful to clarify the basic framework. Then, using this established framework, a second, more efficient approach is presented, and the corresponding results obtained for the distribution of the time constants and for the number of defects in devices is presented. The validity of these results is reinforced though the realization of an aging test in different conditions to those used for the construction of the distribution, the results of which are correctly predicted by the previously obtained parameters. Then, using the information about the amplitude distributions obtained in Chapter 3, and the information about the time constant distribution and number of defects obtained in Chapter 4, a simulation tool for RTN, another for aging phenomena, and an emulator of RTN experimental results, are presented. The characteristics of these tools, as well as their utility, are explained in detail in the corresponding Sections.

In Chapter 5, the focus is shifted to the exploitation of TZV for security applications. In particular, a method that aims the improvement of the reliability of SRAM PUFs is presented. This method, the MTSV method, works by evaluating which cells will have a stronger tendency to always power up to the same value, even when factors such as temperature and voltage variations, or circuit degradation, are considered. The method proves to outstand in all those cases.

Finally, to close this Thesis, Chapter 6 presents the main conclusions of this work.

Chapter 2

Variability characterization framework

Fig. 2.1 shows a block diagram for the development of RAD solutions, including a simple taxonomy of each stage and the different areas of expertise involved in them. In it, it can be seen how a combined effort from expertise in different areas is fundamental. In fact, each stage of this flow is necessary to achieve the final goal of designing circuits that are resilient against variability phenomena. Consider for instance the characterization phase, which is the main focus of this chapter. If a given TDV phenomenon cannot be experimentally characterized, it will not be possible to construct a model that accurately predicts how this phenomenon can affect circuit reliability under different conditions. Therefore, it will not be possible to develop simulation tools to study the impact of this phenomenon in circuit design, and it will not be possible to design circuits that are resilient to that given TDV phenomenon. On the other hand, Fig. 2.1 shows that, at each step of the flow that leads to the development of RAD solutions, there are a number of different choices to be made. Interestingly, the options taken in the different steps are interconnected. Consider, for instance, the work presented in this Thesis. The model used for TDV, the PDO model, is defect-centric, and it can be embedded in stochastic simulation tools. To account for such stochasticity, a large number of transistors must be characterized in different conditions, as will be explained in the next Sections of this Chapter. This massive characterization could in principle be performed following a wafer- or an array-based approach. The latter has been chosen in this Thesis, since it allows a higher density of devices per area and does not require of a dedicated probe station for the measurements.

In this Chapter, a complete framework for the characterization of TDV phenomena at the device level is presented. To this end, first the characterization strategy for each of these phenomena will be described. Then, a number of requirements imposed by that characterization strategy will be listed, and a chip and measurement setup that address those requirements will be presented. After that, a software toolbox for the automated analysis and parameter extraction from the data generated during the characterization tests will be introduced, and some examples of analysis performed with that toolbox will be shown.

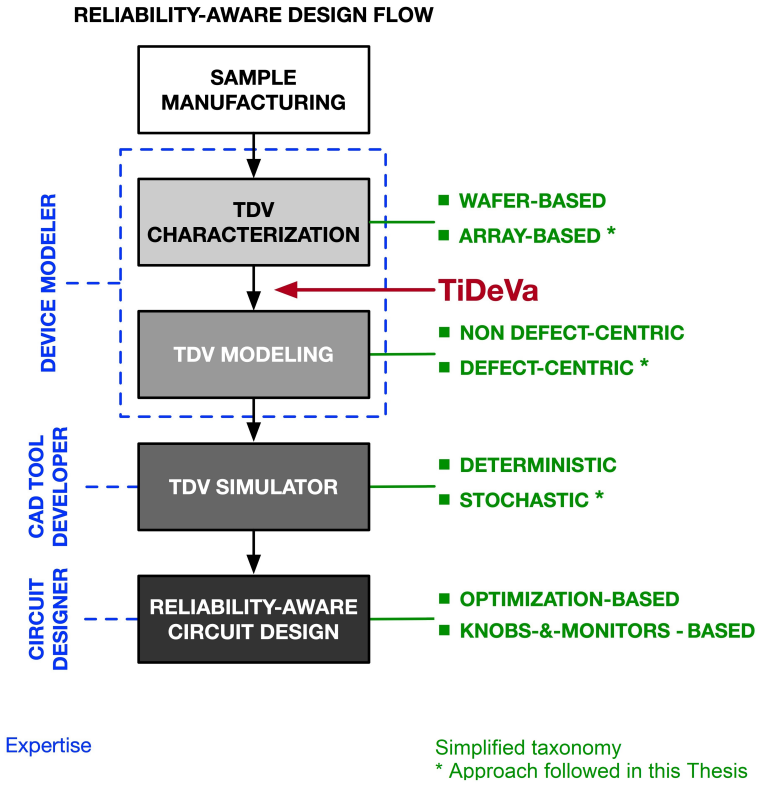

Figure 2.1: Block diagram for the development of RAD solutions, indicating a simple taxonomy of each stage and the different areas of expertise that RAD involves. TiDeVa is the tool developed for the automated analysis of TDV tests and will be presented in Section 2.5.

2.1 Strategy for the characterization of TDV at the device level

The various TDV variability phenomena that have been studied in this work require different types of tests from which the TDV parameters can later be extracted. These tests are presented in more detail below. Before that, there is an important observation to be made. The construction of the TDV model occurs in two distinct phases. First, the defect-level parameters are directly extracted from the TDV tests. This includes information such as the times that emission/capture events take for each individual defect, or the amplitude of the current shift associated to each given defect. Then, in a second stage, those defect-level parameters corresponding to individual defects are used to attain the statistical distributions followed by those defect-level parameters. In turn, those distributions will be described by distribution-level parameters (e.g., a normal distribution would be described by its mean value and standard deviation). This Chapter leads with the extraction of the defect-level parameters from the characterization tests, while Chapters 3 and 4 tackle the construction of the distributions using those defect-level parameters. In particular, Chapter 3 focuses on the analysis of the RTN tests, and uses the

defect-level parameters related to the amplitude of the current shifts associated to each defect to build the corresponding distributions. Then, Chapter 4 shifts to the aging tests and, by using the extracted emission times of individual defects, tackles the construction of the time constant distribution of the PDO model, and the retrieval of the total number of defects in devices. Combining all this information (i.e., distribution of the amplitudes associated to the defects, of the time constants, and total number of defects), it is possible to model and consequently simulate the trapping/detrapping behavior that characterizes the TDV phenomena studied in this Thesis.

2.1.1 Random Telegraph Noise

To measure RTN, each terminal of a transistor must be kept at a given voltage while the drain current of the device is measured. This must be done for thousands of devices and under different conditions, e.g., at different values of V_{GS} and V_{DS}. An example of a real RTN trace measured in this manner is shown in Fig. 2.2. This example corresponds to a simple case in which the current trace displays only one detectable RTN defect, which causes its current to alternate between two levels. The main parameters of that RTN defect (i.e., amplitude of the transition and times-to-emission/capture) are indicated in the figure. Notice that, as indicated in Section 1.3.1, the emission (τ_e) and capture (τ_c) time constants of a defect can be retrieved by averaging the corresponding times-to-emission (t_e) and times-to-capture (t_c). These time constants and the transition amplitudes, together with the total number of defects, are the parameters that can be extracted from the RTN tests, along with their dependence on the operation conditions. In this Thesis, the data generated during the RTN tests have been used mainly to extract the distribution of the current shift amplitudes associated to the RTN defects, and the number of active RTN defects (defects that undergo at least one trapping/detrapping event during the experimental time window). The other defect parameters (i.e., time constants and total number of defects in devices) have been tackled through the aging characterization tests. There are a couple of reasons for this. First, it must be noted that RTN tests must be performed serially (measuring only one device at each given instant), since only one current can be measured at a time with the experimental setup used in this work. Considering that hundreds of devices are measured in a typical RTN test, the characterization times for RTN are not excessively long, and are typically within some tens of seconds, since longer times would lead to unfeasible experimental times. Therefore, these tests are not optimal for the characterization of the distribution of the time constants or the total number of defects in devices, since the first one is expected to span across several decades in the time scale [96], and only a small fraction of the latter one is expected to appear in these tests. Secondly, RTN tests are performed only at nominal voltage values, thus limiting the range of experimental bias conditions. Since the time constant distributions are expected to vary greatly with the bias conditions, the ability to work at very different bias conditions is an interesting option for exploring different regions of such distribution. For these two reasons, the aging characterization tests will be used in this Thesis to tackle this problem, since i) they can be performed using over-the-nominal voltage values, which allows exploring larger regions of the

time constant distribution, and ii) the stress applied in this type of tests can be applied in parallel to a high number of devices, which allows to achieve longer stress times for each device without excessively increasing the test time, and therefore allows to investigate defects with longer capture times.

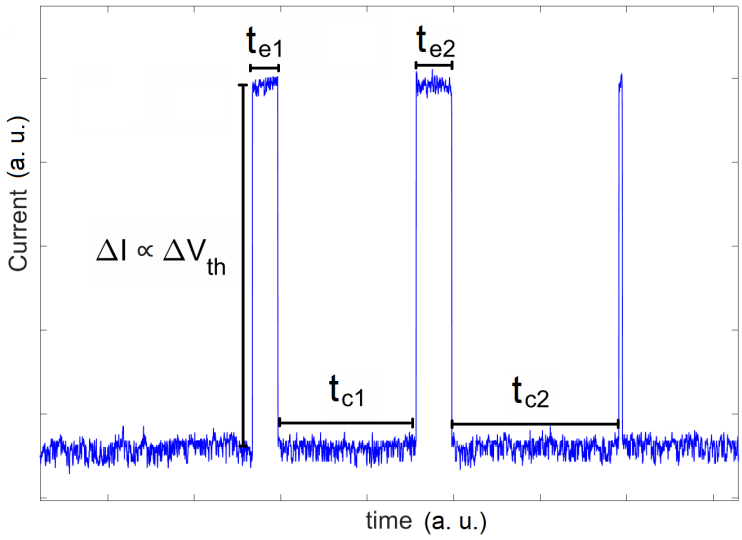

Figure 2.2: Current trace displaying RTN events measured with the chip used in this work and the characterization setup presented in this Chapter. The RTN parameters have been indicated.

2.1.2 Aging phenomena

In the case of the aging phenomena (i.e., BTI and HCI), the characterization tests will be performed in a conventional Measurement-Stress-Measurement (MSM) approach [97], [44]. First, each pristine device must be characterized before any stress has been applied to it to account for TZV. This "fresh" characterization is done by means of a drain current vs. gate voltage (I_{DS}-V_{GS}) measurement and will serve as a reference for each device. After this, consecutive stress-recovery cycles will be applied to each transistor. During the stress phase, $|V_{GS}|$ is kept at a high voltage (between 1.2V, the nominal voltage of the technology, and 2.5V) and $|V_{DS}|$ at 0V for the BTI tests, while for the HCI tests both $|V_{GS}|$ and $|V_{DS}|$ are kept at a high value. The objective of applying these high, over-the-nominal voltages is to perform accelerated aging, and to emulate years of degradation under nominal operation conditions in much shorter times. In the tests presented in this work, 5 stress-recovery cycles have been performed for each device. The duration of the stress phases within these tests increases exponentially (1s, 10s, 100s, 1,000s and 10,000s). On the other hand, both $|V_{GS}|$ and $|V_{DS}|$ are kept at low voltages (i.e., below 1.2V) during the recovery phases, the duration of which has been kept constant at 100s. Notice that, if no parallelization is employed and the MSM cycles are applied serially to the devices, this would lead to extremely long measurement times. Consider for illustration the above mentioned MSM test with 5 stress and 5 measurement cycles. The characterization of each device would take 1s + 100s + 10s + 100s + 100s + 100s + 1,000s + 100s + 10,000s + 100s $\approx$ 11,600s, which is roughly 3 hours. Therefore,

if thousands of devices must be measured at different conditions, this would lead to months or years of continued testing. To avoid this, a smart parallelization scheme should be implemented, in which a large number of transistors can be stressed simultaneously while one of them undergoes a recovery phase, during which its drain current is measured. Then, the discrete current jumps during the recovery phase caused by charge detrapping events can be recorded. Notice that this strategy does not allow the monitoring of capture events during the stress phases, since several devices are stressed in parallel in these phases and their currents are not measured in the meanwhile. However, the influence of the capture times is accounted for through the stress phases with different times and through the application of different stress voltages. The fact that capture times are expected to be distributed in a logarithmic time scale [96], [98], [99] is the reason for the utilization of exponentially increasing stress phases. Furthermore, even if the capture events during the stress phase were recorded, it would not be possible to transform the corresponding current shifts into threshold voltage shifts, since the transistor models provided by semiconductor foundries are not valid at such high biasing voltages.

Then, the direct goal from these tests will be the extraction of the emission events produced during the recovery phases. In particular, both the amplitude of the current shifts (or, alternatively, of the threshold voltage shifts) and the emission times of those emission events will be recorded. These emission times will be used to construct the corresponding time constant distribution of the defects, as will be seen in Chapter 4. Two examples of BTI recovery traces measured in this work are shown in Fig. 2.3. In the trace at the top, only BTI recovery events are present. However, the analysis of the recovery traces can often become more complex due to the coexistence of such detrapping events with RTN events, as can be seen in the bottom trace.

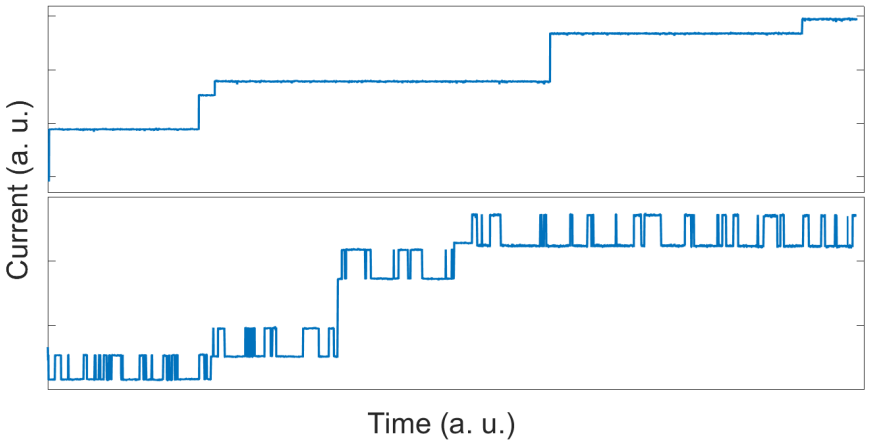

Figure 2.3: BTI recovery traces measured with the Endurance chip and the setup presented in this work. At the top, only BTI detrapping events are present. At the bottom, the BTI detrapping events are mixed with RTN transitions. The coexistence of such differents types of behavior can make the parameter extraction challenging.

2.2 Requirements for the characterization of TDV at the device level

In this Section, the main requirements that the different elements involved in the characterization process should fulfill, according to the characterization strategy outlined above, are listed:

- Due to the stochasticity that TDV phenomena display in transistors in the nanometer range, and to the necessity of studying different phenomena under different bias conditions and temperatures, a large number of devices should be characterized so that the results obtained for each of these phenomena are statistically significant. Additionally, both PMOS and NMOS transistors of different sizes should be included.
- Both the characterization chip and setup should allow the characterization of the different TDV phenomena, namely RTN, BTI and HCI, together with TZV.
- The characterization system should enable individual access to each device, together with the accurate application of the needed voltages. This can pose a challenge in array-based chips such as the one used in this work, which can present undesired voltage drops. Overcoming this problem is fundamental, since TDV phenomena are highly bias-dependent.
- The characterization system should allow the accurate application of temperature within a broad range, since TDV phenomena display a strong dependence on temperature.
- It should be possible to perform a smart parallelization scheme for the aging (BTI and HCI) tests. Such a scheme would allow to reduce in several orders of magnitude (i.e., from months or years to hours or days) the time required to perform these tests as compared to an approach with no parallelization.
- A customized software that allows the automated generation, launching and control of the experiments is fundamental. The alternatives (manual definition and launching of the experiments, or individual input of the commands that control the instrumentation) become unattainable when a large number of devices must be measured at varying operation conditions and with a known and precise timing.
- Considering the necessity for statistically-significant characterization of the different TDV phenomena, a massive amount of experimental data will be generated during the characterization tests. To account for this, a software tool for the automated analysis of these data can enormously facilitate the extraction of the relevant TDV parameters. In the absence of such a tool, this procedure will be not only tedious, but also extremely time consuming.

All these requirements are addressed in a combined manner by the different elements that compose the characterization system used in this work. These are described,

together with some examples of the TDV results obtained with them, in the next Sections of this Chapter.

2.3 Endurance: a transistor array for variability characterization at the device level

The chip used for the device-level characterization experiments presented in this work, named Endurance, has been introduced in [95]. It is the first design that allows the accurate and statistically significant characterization of TZV, BTI, HCI and RTN.

The Endurance chip has been fabricated in a 1.2-V, 65-nm commercial CMOS technology, and has a chip area of 1.8 mm $\times$ 1.8 mm. It contains 3,136 MOS devices, half of which are PMOS and half of which are NMOS, divided into 4 submatrices, each one containing 784 devices. This high number of devices per chip sample allows a statistically significant characterization of the TDV phenomena, which is necessary to account for their stochastic nature at this technology node. Each device is embedded into a unit cell that includes the necessary circuitry to allow, together with a full custom digital control circuitry (common to the rest of the array), individual access to the device. Additionally, each of these unit cells incorporates a Force-&-Sense architecture that ensures the accurate application of the desired voltages through the compensation of any possible voltage drops. The Force-&-Sense system (that comprises both the part within the chip, and another one outside of the chip corresponding to the experimental setup, as will be seen in the next Section) is a type of feedback circuit, in which the Sense path is connected to the Force path at a point close to the device terminal, called the sensing point. Then, the Sense can evaluate which is the actual voltage value that is being applied, which will be generally lower than the one set by the Force path due to undesired voltage drops. Then, the output value of the Force path is varied until the Sense path senses the correct voltage value.

Another essential feature of the Endurance chip is its ability to enable a smart parallelization scheme for aging (i.e., BTI and HCI) experiments, which significantly reduces the experimental time of these tests. To this end, both the drain and gate terminals of each device have separate connections to a measurement path and to a stress path. Then, several devices can be stressed in parallel by connecting their gate and drain terminals to the corresponding stress paths while, simultaneously, a single device is measured by connecting its terminals to the corresponding measurement paths. The digital circuitry mentioned above is used to control to which paths each device is connected to. To illustrate the importance that this parallelization scheme can have, one can consider a typical aging experiment in which 784 devices (a complete submatrix of the Endurance chip) are characterized with 5 stress-recovery cycles, with exponentially increasing stress times of 1s, 10s, 100s, 1,000s and 10,000s, and recovery times of 100s for all cycles. Such an experiment would require 104 days if no parallelization scheme is implemented. On the other hand, with the parallelization scheme allowed by the Endurance chip and used in this work, only 4

days are necessary [100], [101].

The different features enumerated in this Section, such as the on-chip Force-&-Sense architecture or the possibility to parallelize the stress tests by having duplicated paths for both the gate and drain terminals, are complemented by the characteristics of the experimental setup, which will be described in the next Section.

2.4 Setup for the characterization of variability at the device level

Fig. 2.4 depicts the main features of the experimental setup used in the characterization tests presented in this work, which has been presented in [101]. These are:

- A fully customized printed circuit board (PCB), which includes shielded triaxial connectors to allow the access to the analog signal chip pads. It includes connectors for both the Force and the Sense paths of the chip.
- An Agilent E3631A power supply for the biasing of the PCB and the chip.
- A Keysight B1500 Semiconductor Parameter Analyzer (SPA). This instrument incorporates 4 High Resolution Sense Measurement Units (HRSMU), each of which includes Force-&-Sense triaxial outputs for accurate voltage application and current measurement.
- A T-2650BV Thermonics temperature system, which allows the accurate application of temperatures in the range between -40^{o}C and 170^{o}C.
- A USB Data Acquisition System (DAQ) from National Instruments, which provides the digital signals that control the chip.
- A IEEE 488.1 GPIB BUS is used for the communication between the controller (a laptop computer) and the different instruments.

In a typical variability characterization test in which hundreds of devices are measured, often at different bias conditions, and in which a precise timing is fundamental (e.g., a BTI test involving several stress-recovery cycles in which all devices must experience exactly the same stress and recovery phases), thousands of GPIB commands must be generated and sent to the instrumentation. To facilitate this otherwise unattainable task, a software toolbox for the automation of the characterization tests is used. This toolbox, which has been presented in [101], works under the Matlab® programming environment, and allows the user to define the desired characterization tests in just a few seconds. Then, it sends the corresponding GPIB commands to the instrumentation at the precise required timing, with no user supervision required. Additionally, this tool manages the Data Acquisition System, which in turn sends the signals in charge of controlling the digital circuitry of the

chip, necessary for the selection of the unit cells and their corresponding operation modes, in this way connecting the terminals of the desired devices to the corresponding analog paths for the application of bias voltages and the measurement of currents.

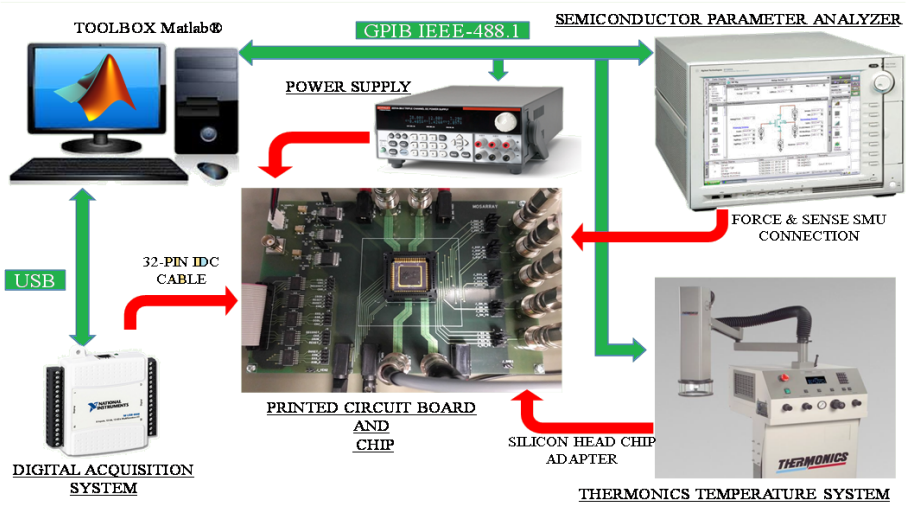

Figure 2.4: Schematic representation of the experimental setup used in this work.

2.5 TiDeVa: a toolbox for the automated analysis of Time-Dependent Variability

The stochastic nature of TDV phenomena in technologies in the nanometer range does not only bring along requirements in terms of circuit design or hardware setup. It has already been said that thousands of devices have been characterized under very different conditions, such as current measurements with constant V_{GS} and V_{DS} in the case of RTN tests (at various V_{GS} and V_{DS} values), or current measurements during the recovery phases of several stress-recovery cycles in the case of aging tests. This leads to the generation of enormous amounts of experimental data, i.e., thousands of current traces with a very broad variety of features such as noise levels or number of trapping/detrapping events, among others. Therefore, the manual analysis of all these data and the consequent TDV parameter extraction would become an unattainable task even for the most skilled user. For this reason, developing software tools devoted to the automated and accurate analysis of the TDV characterization data becomes as an important feature within the characterization process as the design of the characterization circuit or setup. To this end, the TiDeVa toolbox was introduced in [102]. This software tool can perform, in a fully-automated manner (i.e., with no supervision from the user) the analysis of the huge amounts of data generated during the TDV tests, and the extraction of the necessary parameters to construct the appropriate models for the TDV phenomena. The engine of TiDeVa makes use of a method based on the Maximum Likelihood Estimation (MLE) technique [103], which will be explained in greater detail in Subsection 2.5.2, to analyze the current traces generated during the TDV characterization tests and extract from them all the information about the trapping/detrapping events needed

for a defect-centric model such as the PDO. Furthermore, TiDeVa is equipped with an intuitive and user-friendly Graphical User Interface (GUI) to help the user launch the desired analysis. The GUI of this toolbox also allows the user to visualize both the processed current traces, and the statistical results for the extracted parameters, such as the distributions for the defect amplitudes or their time constants. Fig. 2.5 shows some tabs of TiDeVa's GUI.

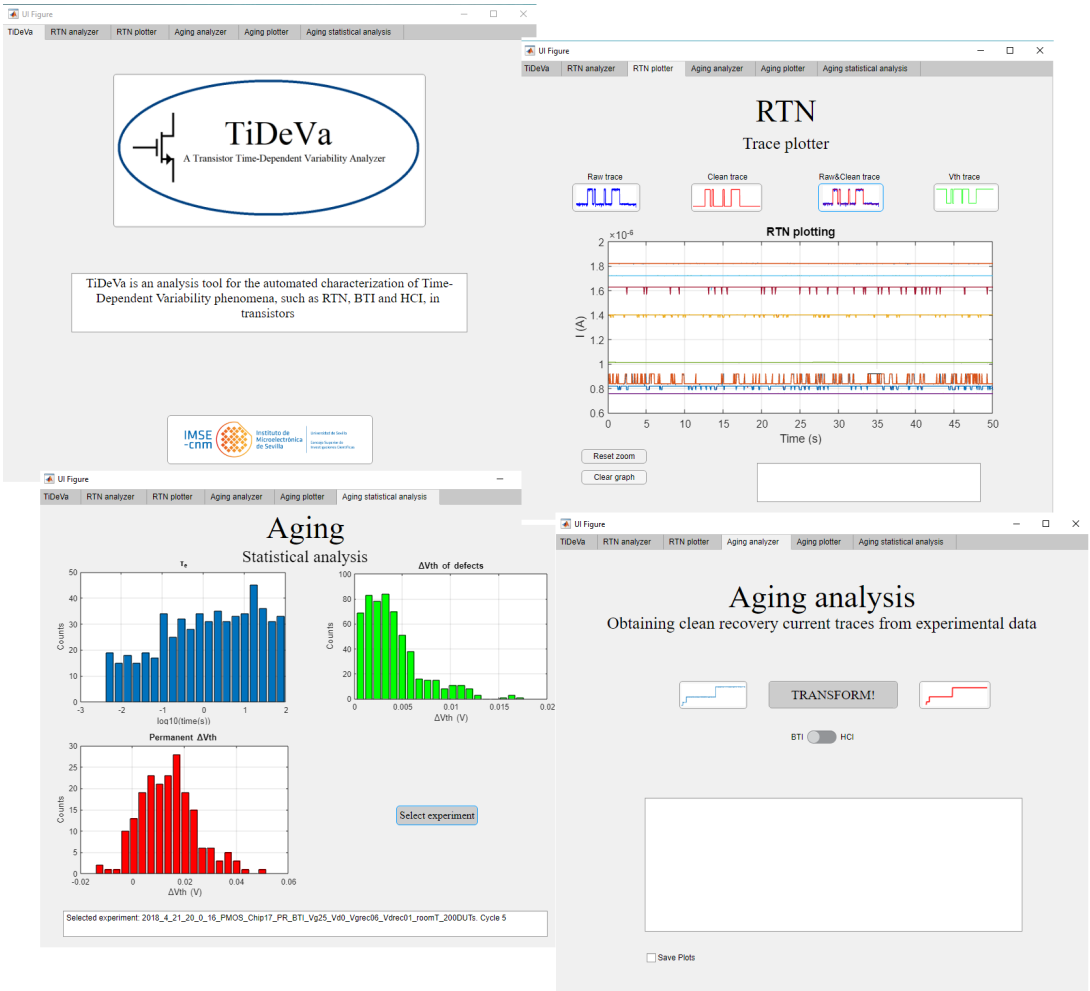

Figure 2.5: Some tabs of the TiDeVa toolbox. Clockwise, starting from the top-left corner: i) initial tab, ii) tab for the graphical visualization of the RTN current traces, iii) tab for the analysis of aging experiments data, and iv) tab for the visualization of the statistical distributions obtained for the parameters of an aging experiment.

Before approaching the description of the MLE method that TiDeVa uses for the analysis of TDV tests, it can be useful to briefly discussed other methods that aim at the extraction of the parameters that characterize the charge trapping/detrapping in/from defects from measured current traces.

2.5.1 Existing methods for the TDV defect parameter extraction

In the last years, the increasing concern about the impact of TDV on deeply scaled technologies has led to the development of a number of TDV analysis methods. Since TDV phenomena reveals a stochastic nature in these technologies, the characterization of a large number of devices is necessary to obtain statistically significant information about it. Therefore, those analysis methods must be not only accurate, but also automated, to perform the parameter extraction in a feasible time and without the supervision from the user. Additionally, since these phenomena display a discrete nature due to the fact that they are caused by the trapping/detrapping of individual charge carriers in/from defects in the devices, many of these methods are based on the detection and analysis of the discrete current shifts caused by those trapping/detrapping events. In this Subsection, the focus will be set on such methods.

Several reported methods to extract the RTN parameters are based on the Time Lag Plot (TLP) technique or its derivatives [33]. Given a current trace, its TLP is constructed by plotting the i-th current point in the x-axis and the $(i+1)$-th point in the y-axis for the complete trace. Then, the points in the diagonal of the TLP (consecutive points with the same current value) correspond to the RTN current levels. On the other hand, the off-diagonal data points correspond to RTN transitions between different current levels. Then, parameters such as the RTN transition amplitudes could in principle be extracted from these current levels. However, the identification of these current levels may be hindered by the existence of background noise due to other physical processes, as well as the noise introduced by the measurement circuitry and the equipment itself, or by the coexistence of a large number of RTN defects. Fig. 2.6 displays two current traces that contain RTN, together with their corresponding TLPs. While Fig. 2.6 (a) corresponds to a trace in which only one RTN defect is clearly detectable, and the corresponding TLP is easy to analyze, Fig. 2.6 (b) corresponds to a trace in which there are several detectable RTN defects, which translates into a more complex TLP.

To overcome this limitation, an improved version of the TLP, the weighted TLP (wTLP) was introduced [104], [105]. For each point of the TLP with coordinates (I_i, I_{i+1}), a bivariate normal distribution is defined:

$$\phi_i(x,y) = \frac{1}{2\pi\alpha^2}\left(\frac{-[(I_i - x)^2 + (I_{i+1} - y)^2]}{2\alpha^2}\right) \tag{2.1}$$

where α represents the standard deviation of the background noise. This distribution symbolizes the probability that the point (I_i, I_{i+1}) corresponds to a current level or to an RTN transition in the location (x, y) of the TLP space.

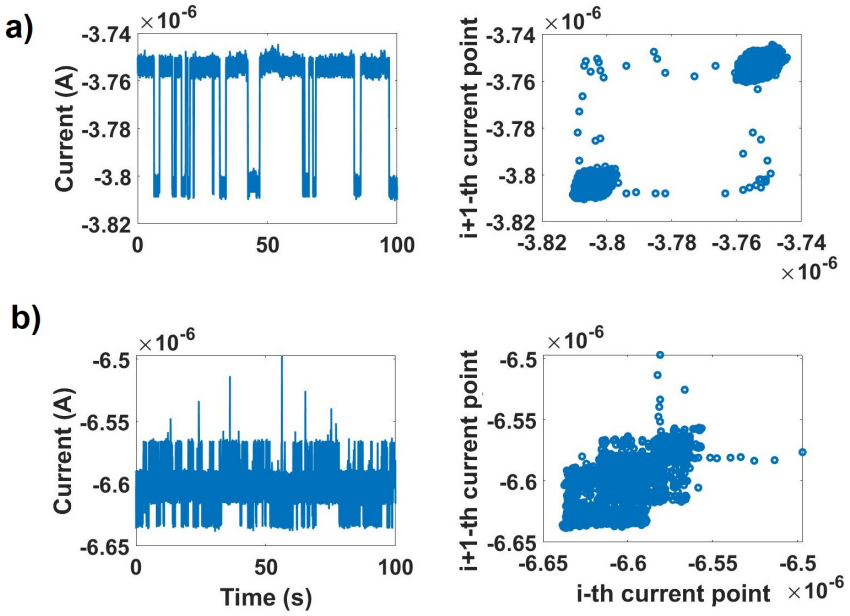

Figure 2.6: Current trace and the corresponding TLP for a device with one detectable RTN defect (a) and a device with several RTN defects (b).

.

Then, the weighted time lag function Ψ can be defined as:

$$\Psi(x, y) = K \sum_{i=1}^{N-1} \phi_i \tag{2.2}$$

where K is a normalization factor that ensures that the maximum value of Ψ is equal to 1, and N is the number of current data points. In this way, the contribution of each point of the TLP to Ψ is weighted by the distance between the position of this point and (x,y), so that Ψ takes higher values in the most populated regions of the wTLP. This technique generates a very visual result, which is adequate for human inspection of the intricacies of the RTN transitions. Fig. 2.7 displays an example of complex TLP, together with the corresponding wTLP, which provides a clearer visualization.

However, a more quantitative approach is needed for an automated extraction of the numerical parameters that characterize RTN. To this end, it is possible to build a histogram of the current data points, and fit it with the diagonal of the obtained Ψ. This fitting is performed by varying the standard deviation α. The best match between the Ψ diagonal and the current histogram should be obtained when α equals the standard deviation of the experimental background noise. Then, by extracting the maxima of the fitted diagonal function, it should be possible to obtain the location of the current levels, and, thus, the amplitude of the RTN transitions. Furthermore, the wTLP method has also been applied to the analysis of BTI recovery current traces [106], [107]. However, as shown in [103], the choice of the bin size of the current histogram in this method can yield different positions of the current

levels, or even a different overall number of current levels, hence determining an incorrect number of RTN or BTI defects, or the amplitude of the transitions.

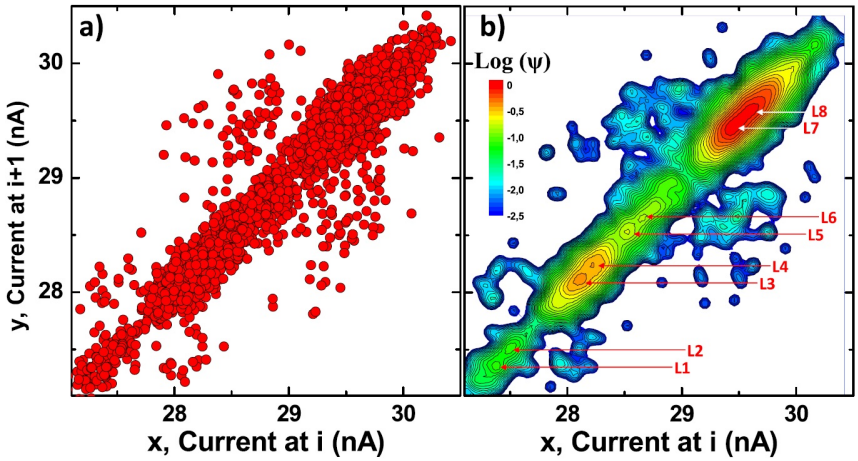

Figure 2.7: TLP of a current trace that displays RTN (a), together with the corresponding wTLP (b). Taken from [104]

.

To overcome this ambiguity in the selection of the number of histogram bins, the approach presented in [30] can be followed. In it, an enhanced TLP (eTLP) is constructed to assess how often each point of the TLP is occupied. This is done by building a two-dimensional histogram with a bin size equal to the minimum resolution of the experimental equipment. Then, the maxima along the eTLP histogram are taken as the current levels. However, since each bin size is equal to the experimental resolution, consecutive current points that do not correspond to exactly the same current value are not taken into account. Therefore, since any real experiment presents a certain level of noise, an enormous amount of data points will be necessary to obtain good statistics. This would result in much longer acquisition and computational times.

2.5.2 TDV parameter extraction using a Maximum Likelihood Estimation method

To overcome the limitations of the methods presented in the previous Section, a new approach to extract the parameters from current traces obtained in TDV tests will be developed in this Thesis. This is a method based on the Maximum Likelihood Estimation (MLE) technique [108], [109], and composes the engine of the TiDeVa toolbox for the TDV parameter extraction that was introduced in Section 2.5 [102]. A block diagram depicting the main steps of this method is shown in Fig. 2.8. These steps, which will be later described in greater detail, are:

1. Detection of the M current levels through a MLE-based optimization process.
2. Elaboration of a clean, noise-free current trace.

3. Identification of each transition between two different current levels corresponding to a charge trapping/detrapping event in/from a defect.
4. Extraction of the defect parameters.

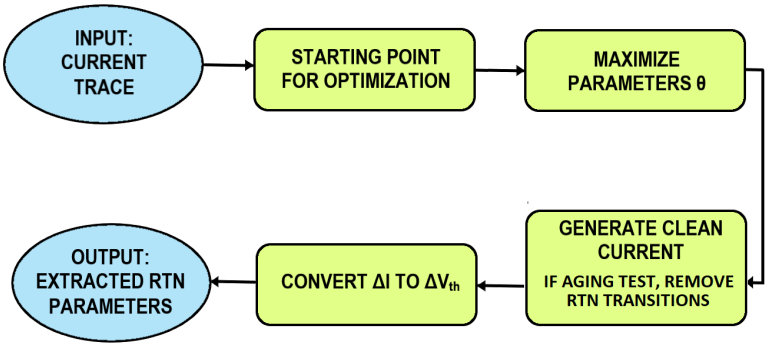

Figure 2.8: Flow diagram of the main steps of the MLE-based algorithm used to extract the RTN parameters.

.

2.5.2.1 MLE-based detection of the current levels

Let us consider a device with a number of defects that undergo the trapping/detrapping of charge carriers that originate TDV phenomena. Because of the trapping/detrapping events of these defects, the measured current trace will have M discrete current levels. The background noise can be approximated by a Gaussian distribution which is independent of the different current levels. Then, the measured current trace can be considered as samples of a probability density function formed by the superposition of different Gaussian distributions corresponding to the different current levels:

$$f(I|\theta) = \frac{1}{K\sqrt{2\pi\sigma^2}} \sum_{j=1}^{M} A_j e^{-\frac{(I-I_{Lj})^2}{2\sigma^2}} \tag{2.3}$$

where $\theta = \{\sigma, A_1, ..., A_M, I_{L1}, ..., I_{LM}\}$ is the vector of parameters for this probability density function. In particular, σ represents the standard deviation of the background noise, I_{Lj} is the value of each current level, and A_j is its height in the probability density function. K is a normalization factor which ensures that the area below the probability density function is unity:

$$K = \sum_{j=1}^{M} A_j \tag{2.4}$$

Then, N samples $\{I_1, ..., I_N\}$ of the current trace can be considered. It can be assumed that all samples are independent and identically distributed. Therefore, the joint probability density function of all observations can be defined as:

$$f(I_1, ..., I_N|\theta) = \prod_{i=1}^{N} f(I_i|\theta) \tag{2.5}$$

The Maximum Likelihood Estimation is based on the principle that the likelihood of a set of parameter values θ, given the observed results, is equal to the probability of obtaining those results as a function of θ, that is:

$$L(\theta|I_i) = \prod_{i=1}^{N} f(I_i|\theta) \tag{2.6}$$

Having a set of parameters θ, the probability density function describes how probable is to obtain a given current value in a measurement. Since the characterization test already yields those current values, the reversed problem is faced: given the observed data points, it is necessary to obtain the probability density function (and therefore the parameters θ) that is most likely to have produced those data. Then, the most likely parameter values θ given the recorded data can be obtained through the maximization of (2.6). Considering that the number of measured current points N can amount to thousands or even millions, it is more convenient to work with the natural logarithm of (2.6); otherwise, the calculation would easily exceed the ranges of floating point numbers in digital computers. Then, the multiplication in (2.6) can be converted into a summation. This implies no difference in terms of finding the parameters that maximize (2.6) because the logarithm is a monotonically increasing function.

Different optimization algorithms can be applied for this maximization problem. In this work, a deterministic local search method has been used for efficiency reasons. This is the derivative-free simplex search method presented in [110]. The convergence of such a method depends on having an adequate starting point. The location and amplitudes of the peaks of the current histogram are good starting points to this end. To avoid that some local maxima caused by the background noise are incorrectly identified as current levels, a minimum distance between the histogram peaks is required. In this manner, if two maxima of the current histogram are located within a distance considered to be smaller than the background noise, only the largest one is considered for the starting point of the optimization process, and the other one is discarded. When the optimization process converges, the M discrete current levels of the trace are obtained. The extracted M current levels for a measured current trace can be observed in Fig. 2.9.

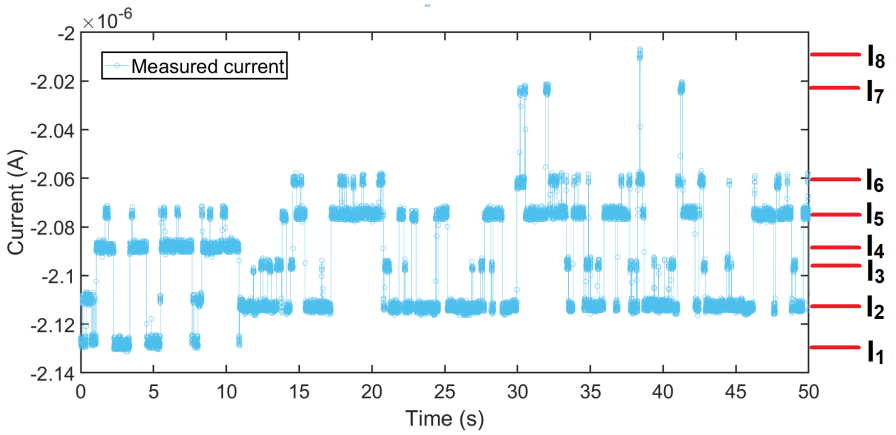

Figure 2.9: Measured current trace for a PMOS device in the Endurance chip displaying RTN, and the corresponding current levels extracted by the MLE method.

.

2.5.2.2 Generation of a clean current trace

Once that the M discrete current levels have been identified, the next step is to generate a clean or background-noise-free current trace from the experimental one. For this, the current value of the closest of the M current levels is assigned to each experimental current point. The generation of such a clean current for a measured trace is shown in Fig. 2.10 for the experimental current trace displayed in Fig. 2.9.

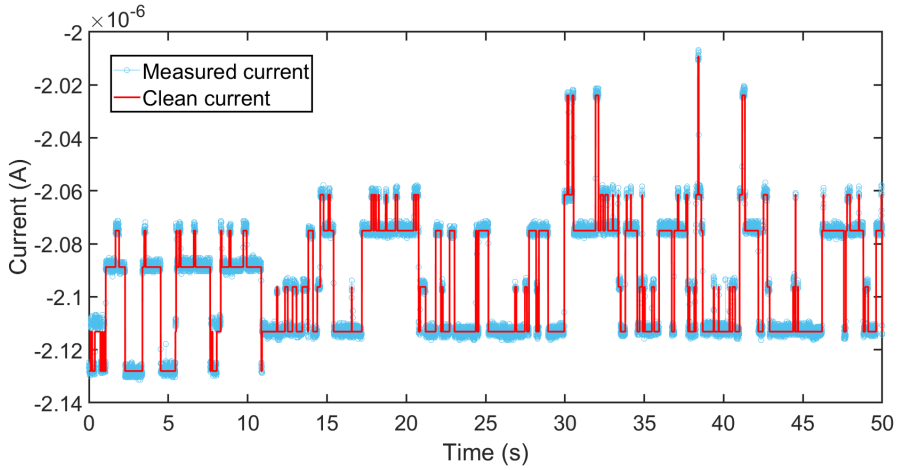

Figure 2.10: Experimentally measured current trace (blue), together with the processed clean current trace (red).

.

2.5.2.3 Identification of the trapping/detrapping transitions

Since the background noise has been removed in step 2 and only the clean current levels remain, the observed current shifts correspond to charge carrier trapping/detrapping transitions. Therefore, it is straightforward to identify these transitions,

and the number of detectable defects in the trace can be determined.

Although the MLE method has been presented as a tool to analyze current traces that display RTN, it can also be used to analyze the current traces obtained during the recovery phases of the aging tests. As explained in the characterization strategy in Section 2.1.2, and as it will be seen in Chapter 4, in the procedure following for the construction of the PDO model, only single emission events (i.e., defects that undergo a single emission event during the time window without a subsequent capture event) are used. Therefore, in the case of aging tests, defects that do not undergo a single emission event (i.e., defects that undergo both trapping and detrapping events, which induce current shifts of equal amplitude but opposite sign) are "removed" from the clean trace, and only recovery detrapping events are left.

2.5.2.4 Parameter extraction

Once the RTN and aging transitions have been identified, the time constants (τ_c, τ_e) and amplitude η that describe the RTN transitions are extracted. These time constants are calculated as the average time-to-capture/time-to-emission for all the capture/emission events of a given defect. In contrast, in the aging tests, where single emission events in which only one detrapping event occurs for each defect during the recovery, the time-to-emission t_e at which the emission event occurs is recorded, and η is directly extracted as the amplitude of that current transition. Additionally, the amplitude of the current shifts of both RTN and aging recovery transitions can be converted into a threshold voltage shift by applying the method presented in [111]. This method can be divided into two main steps: first, determining the fresh transistor parameters of each device and, second, determining the evolution of those parameters during the RTN or aging current measurements. In the following, these two steps are explained in further detail.

The main idea of the first step of the process is to try out different values of some of the parameters of the BSIM model to try to replicate an experimental I_{DS}-V_{GS} curve measured in the fresh device (i.e., before any stress is applied to it). Then, the parameters used for the simulation of the curve that better fits the experimental one can be assigned to that device. The procedure followed to obtain the simulated curves that will be compared to the experimentally measured ones is the following. First, a grid of possible transistor parameter values is generated using foundry-provided TZV information. For instance, μ_0 and V_{th0}, which are the parameters used, are uniformly swept in the range of $\pm 6\sigma$ around their nominal value. This results in a grid with all possible combinations of μ_0 and V_{th0}. In our case, 1,000 values of V_{th0} and 100 values of μ_0 have been used in this process, which yield a grid of 100,000 points. Such a grid is represented in Fig. 2.11. Then, taking each parameter combination from that grid, a I_{DS}-V_{GS} curve is simulated (in our case with HSPICE), which results in a total of 100,000 simulated curves. Then, each of these set of simulated curves is compared to the experimentally measured I_{DS}-V_{GS} curve by the calculation of the Mean Squared Error between them. Then, the simulated curve that yields a smallest MSE is considered to be the best match to the experimental curve, and the transistor parameters used for its simulation are

assigned to the device. In this way, it is possible to obtain the fresh transistor parameters of the device.

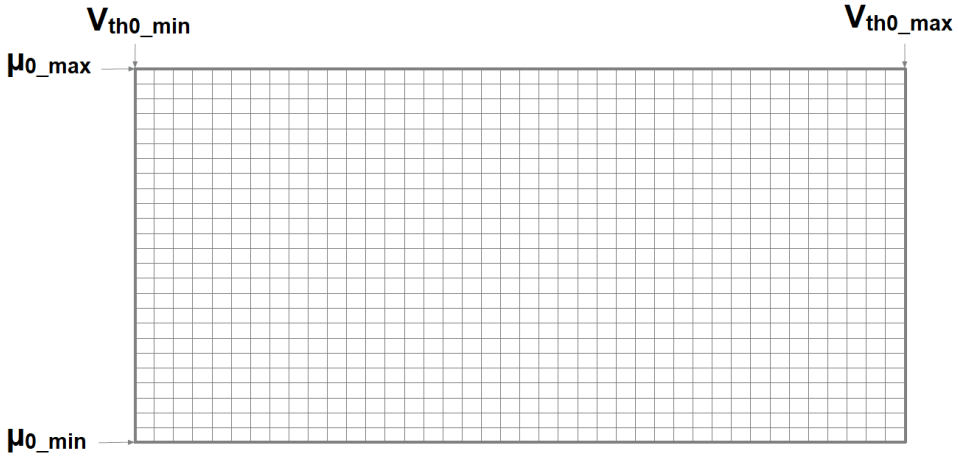

Figure 2.11: Representation of the grid that contains all possible combinations of the μ_0 and $_{th0}$ parameters.

The main goal of the second step of the process is evaluate the evolution of the V_{th0} value during a current measurement, which can be an RTN current measurement, or a recovery cycle in an aging MSM test. For this, a second I_{DS}-V_{GS} is measured at the end of the current measurement, and the transistor parameters are extracted with the procedure explained above. Then, it is assumed that the mobility of the device does not change significantly during the RTN measurement, and that in the case of the aging measurement, although it may degrade due to the application of the stress, it does not recuperate significantly once the stress is removed [111]. This is depicted in Fig. 2.12. Then, since the mobility during the current measurement is assumed to be equal to the one extracted from the I_{DS}-V_{GS} curve performed at the end of the measurement, only the set of simulated curves with the closest mobility to the extracted one in this second I_{DS}-V_{GS} curve are selected among the 100,000. Then, for each point of the current trace, it is possible to evaluate which curve from that set has a closest current value at the experimental bias conditions. Then, since all those curves have the same mobility parameter, the variations in current are assigned to variations in threshold voltage. Thus, the threshold voltage that was used to generate the curve that better fits each current point is assigned to that current point. Then, a threshold voltage has been assigned to each current point in the trace, and therefore a threshold voltage trace is obtained. This allows to transform the current shifts caused by charge trapping/detrapping into threshold voltage shifts.

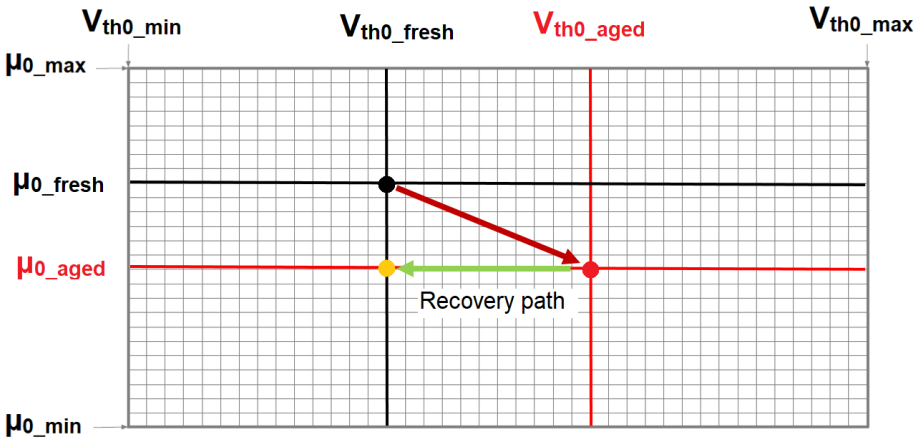

Figure 2.12: Representation of the expected evolution of the transistor parameters in an aging test.

2.5.3 Some examples of TDV characterization at the device level performed by TiDeVa

In this Section, some examples of TDV characterization performed by TiDeVa are shown. In particular, results from RTN tests and aging (i.e., BTI and HCI) tests are presented. In both cases, the focus is first set on the parameter extraction from individual current traces generated during the TDV tests. After that, some statistical results obtained with TiDeVa from large sets of devices are illustrated.

2.5.3.1 RTN characterization examples

Fig. 2.13 displays the current traces measured for a single 80nmx60nm PMOS device from the Endurance chip at $|V_{DS}| = 0.1V$ and various $|V_{GS}|$ values. The TiDeVa toolbox processes the raw experimental data and yields clean traces in which the background noise (including sources such as thermal noise of the devices and noise introduced by the measurement setup) has been removed, and only the RTN transitions remain. Additionally, TiDeVa extracts the RTN parameters (time constants and transition amplitude of each defect) from each measured current trace. Then, the dependence of these parameters on the bias conditions can be studied. In this sense, Fig. 2.14 displays the extracted dependence of the amplitude of the RTN defect in Fig. 2.13 on $|V_{GS}|$.

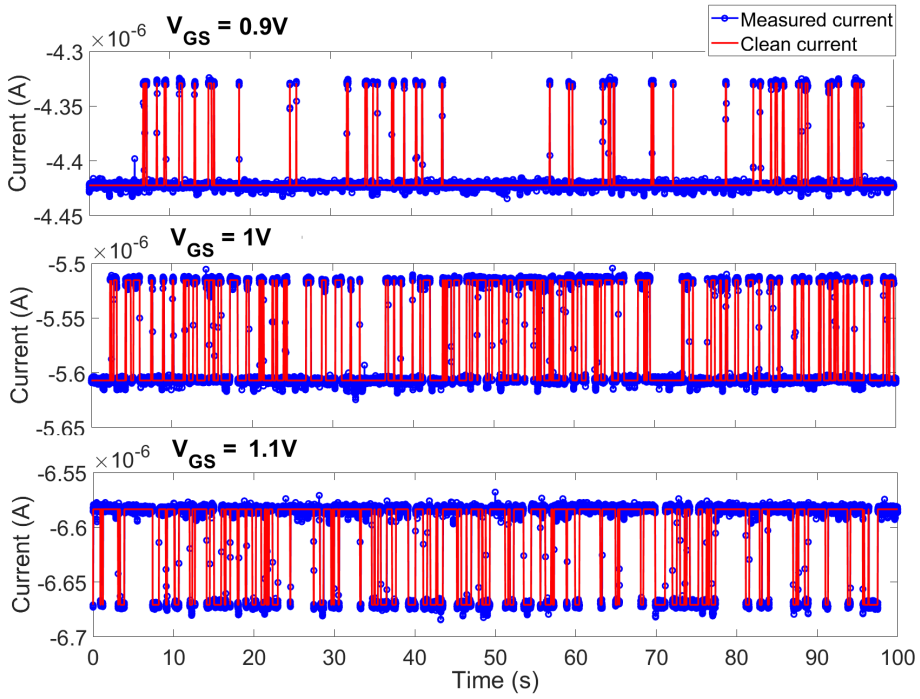

Figure 2.13: Current traces of one 80nmx60nm PMOS device measured at $|V_{DS}| = 0.1V$ and various $|V_{GS}|$ displaying RTN transitions. The blue lines correspond to the raw experimental data, the red ones to the "clean" traces processed by TiDeVa.

.

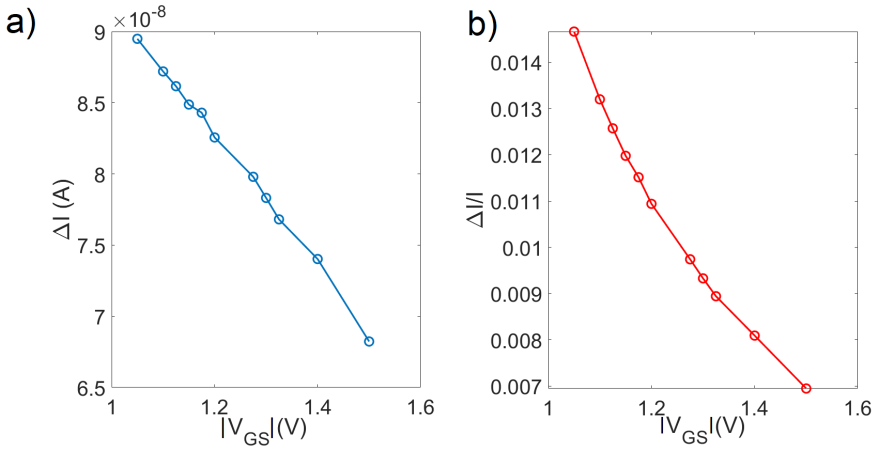

Figure 2.14: Absolute (a) and relative (b) current amplitude of a defect monitored at different gate voltages.

.

Since TDV phenomena reveal a stochastic nature in technologies in the nanometer range, the information extracted from a single transistor is not enough, and the massive characterization of a large number of transistors becomes mandatory. Consequently, Fig. 2.15 displays the statistical distribution of the RTN amplitudes extracted by the TiDeVa software from 500 80nm×60nm PMOS devices measured at various bias conditions. This illustrates the necessity of an automated analysis tool such as the TiDeVa software; manually analyzing hundreds of devices, each of

them at various bias conditions, and extracting from them the necessary parameters to characterize the RTN phenomenon would result extremely burdensome for any user.

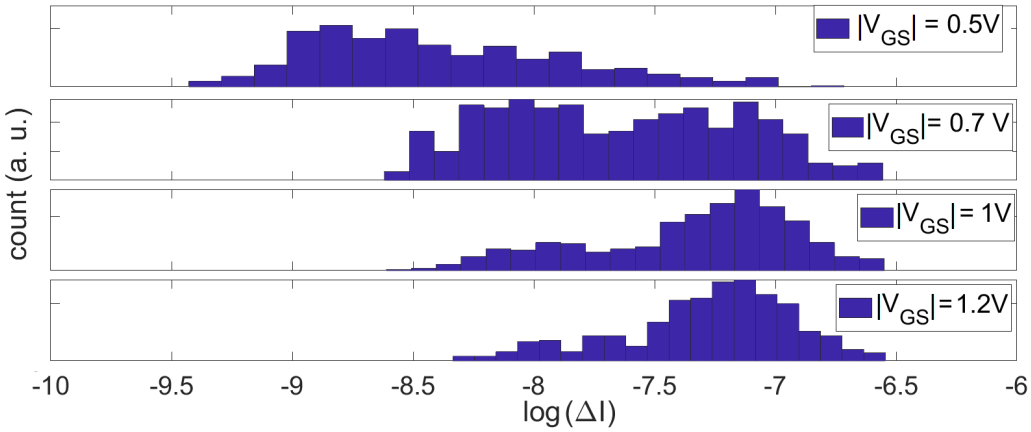

Figure 2.15: Experimental distribution of the amplitude of RTN defects at $|V_{DS}| = 0.1V$ and various $|V_{GS}|$ values. The defects extracted by the TiDeVa tool from 500 transistors have been used to construct these histograms.

2.5.3.2 Aging phenomena characterization examples

Fig. 2.16 displays two examples of current measurements performed during the recovery phases of a BTI test. These two examples illustrate two usual cases: a recovery in which only detrapping events are present (top), and a recovery in which those recovery events are mixed with RTN trapping/detrapping events (bottom). In both cases, the TiDeVa toolbox removes the background noise leaving a clean trace in which only the trapping/detrapping events are present. Then, the software identifies and removes the RTN transitions when present, so that only the recovery detrapping events remain. The reason for this is that, as it will be explained in Chapter 4, only the emission events extracted from the recovery phases will be used during the construction of the TDV model. Once that the clean trace has been obtained, TiDeVa extracts, analogously to the RTN case, both the amplitudes and the emission times of those detrapping events.

As said before, a massive characterization of hundreds of devices must be performed in deeply-scaled technology nodes. Fig. 2.17 displays the cumulative occurrence of the emission times t_e of the defects extracted by TiDeVa at each one of the recovery cycles (after 1s, 10s, 100s, 1,000s and 10,000s of stress) of the BTI test from which the traces in Fig. 2.16 have been extracted. The requirement of an automated tool to avoid the otherwise extremely cumbersome task of this analysis is again highlighted, considering that more than 3,000 devices have been studied in the BTI tests presented in this Thesis, and from them thousands of defects, with their corresponding parameters, have been extracted.

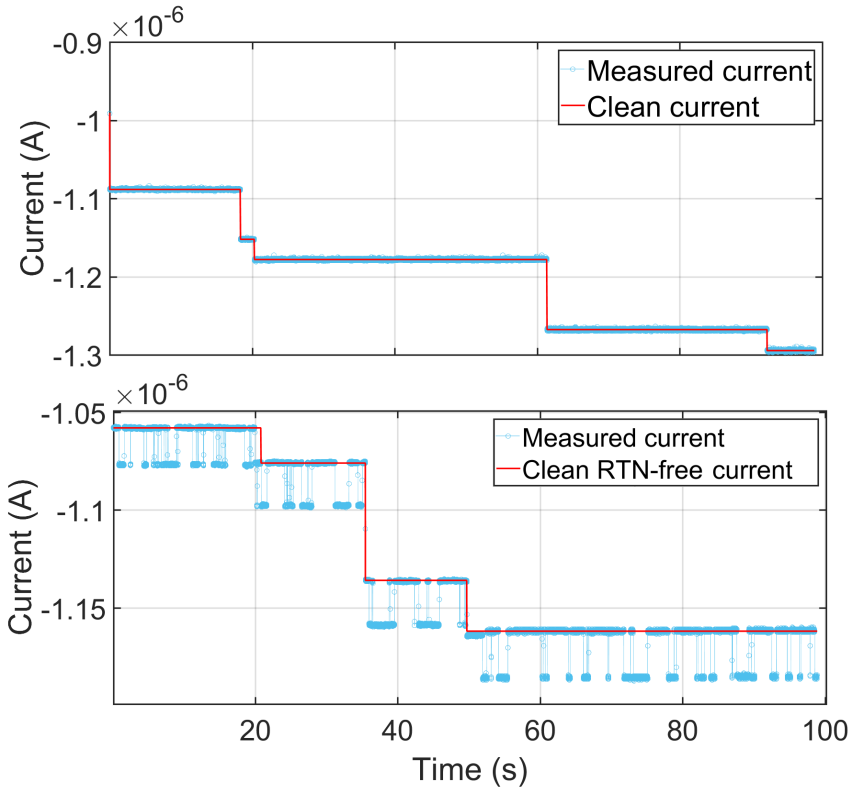

Figure 2.16: BTI recovery traces measured at $|V_{GS}| = 0.6V$ and $|V_{DS}| = 0.1V$ after a stress phase at $|V_{GS}| = 2.5V$ and $|V_{DS}| = 0V$. The processing has been performed with the TiDeVa toolbox.

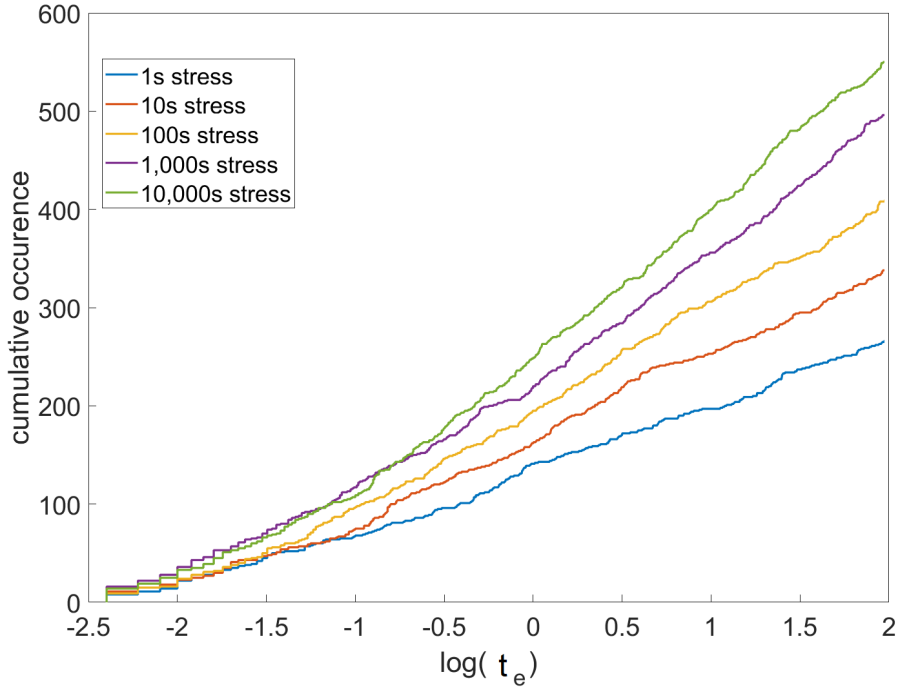

Figure 2.17: Cumulative occurrence of the emission times for each of the five recovery phases in a BTI experiment with $|V_{GS}| = 2.5V$ and $V_{DS} = 0V$ during the stress phase, and $|V_{GS}| = 0.6V$ and $|V_{DS}| = 0.1V$ during the recovery phase.

Chapter 5

Exploiting TZV through the utilization of SRAM PUFs

Chapters 2, 3 and 4 of this Thesis have dealt with the characterization, modeling and simulation of TDV phenomena at the device level. In this Chapter, as it was explained in Chapter 1, the focus is shifted to the exploitation of TZV for security applications. This is done through the concept of PUF and, in particular, of SRAM PUFs, both of which have also been introduced in Chapter 1. The ultimate goal of this part of the Thesis is to develop and experimentally test a method that increases the reliability of SRAM PUFs. This will be done by selecting the SRAM cells that display a more stable power-up response, since the power-up value of SRAM cells is usually used in SRAM PUF responses [76], [77]. Although the power-up stability of SRAM cells is mainly determined by TZV, TDV will also play an important role in this Chapter. The reason for this is that one of the factors that threatens PUF reliability is circuit degradation caused by TDV phenomena [77], [126]. Therefore, the method developed to increase the PUF reliability will have to account for the impact of the previously discussed TDV phenomena.

The structure of this Chapter will be the following one: first, KipT, a chip that includes the SRAM cells used in this work, is presented, and the used experimental setup is briefly described. After that, the main metrics used to evaluate the quality of the PUF, and some existing methods to increase the reliability of SRAM PUFs based on Dark-Bit Masking techniques, are outlined. Following these, the tests performed to evaluate the adequacy of the novel method in order to increase the PUF reliability are explained. These tests include measurements of the pristine circuit at nominal conditions, when temperature and supply voltage variatons and considered, and after the circuit has been degraded in an accelerated manner.

5.1 KipT: an IC for the evaluation of the reliability of SRAM PUFs

This Section presents a chip that can be used to evaluate the adequacy of different methods to increase the reliability of SRAM PUFs, the KipT chip. This chip, which has been designed in a 1.2-V, 65-nm commercial CMOS technology (the same than the one used for the Endurance chip), contains three different main arrays, each one composed of different types of circuits. These are a Ring Oscillator array [127], an array formed by a set of different elementary circuits such as inverters, amplifiers and current mirrors [123], and an array of 6T SRAM cells [128], which will be used to test the novel methodology to increase the reliability of SRAM PUFs and will therefore be the focus in this Section.

5.1.1 The SRAM cell array in the KipT chip

The layout of the SRAM cell array is shown in Fig. 5.1. Each array contains 832 SRAM cells distributed in 32 rows and 26 columns. Following the conventional sizing criteria to achieve good read stability and writeability introduced in Section 1.7, the access and the PMOS pull-up transistors have been sized with $W = 80$nm and $L = 60$nm, while the NMOS pull-down transistors have been sized with $W = 160$nm and $L = 60$nm. Four additional rows on top of the SRAM cell array contain different topologies of sense amplifiers, although these have not been used in the tests, and will therefore not be further discussed.

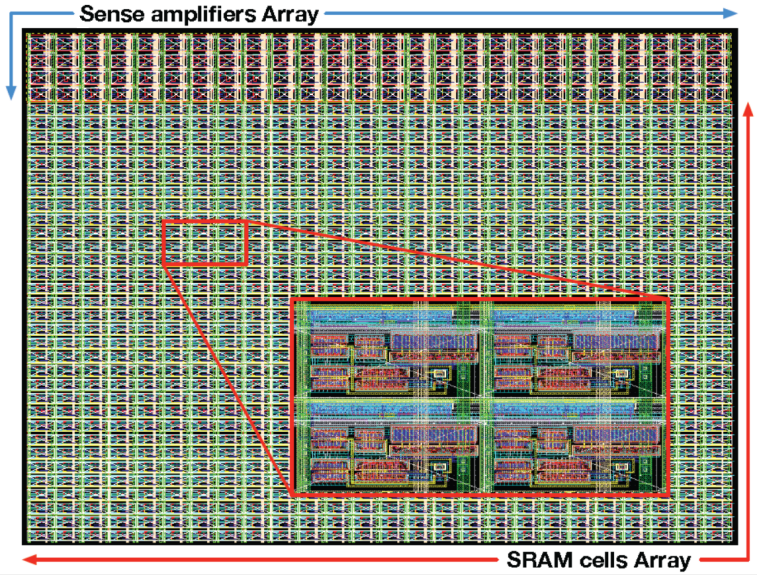

Figure 5.1: Annotated layout of the SRAM array in the KipT chip.

The structure of the SRAM unit cell is analogous in many ways to the one of the Endurance chip [95]. Each SRAM cell is embedded into a unit cell that includes the necessary circuitry to allow, together with a full custom digital control circuitry

(common to the rest of the array), individual access to the different terminals of the given SRAM cell. The array has row and column decoders to individually select each SRAM cell in the array. The input signals to these decoders are provided by two shift registers, which are accessed sequentially through digital I/O pads. For the sake of illustration, the digital signals needed to select one of the cells of the array are displayed in Fig. 5.2. In it, $ICOL$ and $IROW$ correspond to the selection signals that are input serially from the least to the most significant bit, and CKC and CKR to the corresponding clocks. In this case, the selected cell would be the one in the first column, fifth row. The $DECOSET$ signal is sent once that the binary address has been stored in the shift register, in order to send the complete binary selection word to the decoders.

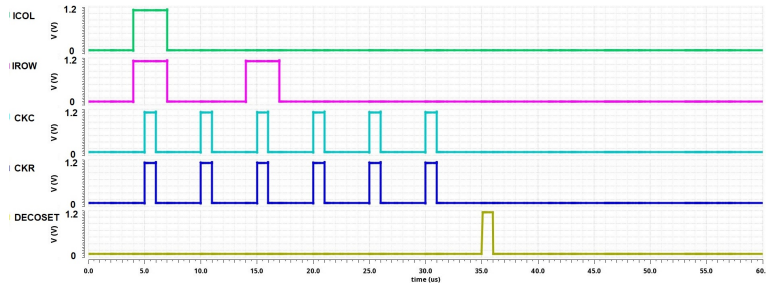

Figure 5.2: Digital signals used to select an SRAM cell of the KipT chip.

The SRAM cell terminals are accessed through a set of analog paths. When necessary, these paths have a Force-&-Sense system to avoid any undesired voltage drop and ensure the accurate characterization of the circuits. As in the Endurance chip, the cells can be set to different operation modes. Depending on the operation mode, each terminal of the cell will be connected to a certain analog path. These operations modes are:

- Measure mode: this operation mode has been conceived to connect the wordline and bitline terminals to the wordline and bitline analog paths, respectively, and the power supply terminals to the nominal power supply path. Under the measure mode, the cell can undergo write and read operations, as well as hold any stored data. The applied voltages can be fully controlled so that any SRAM performance metric can be properly measured.

- Stress hold mode: this operation mode has been designed to connect the power supply terminal of the cell to the stress power supply path. Additionally, the wordline node is connected to the standby path, so that the access transistors are not activated, and the cell remains in hold state. A number of cells can be simultaneously in the stress hold mode, with their power supply terminals connected to the stress power supply path, while another cell is being measured in the measure mode. This feature allows the parallelization of the hold stress. This mode will be used to stress the SRAM cells in order to test the reliability of the PUF against circuit degradation.

- Stress write/read mode: this operation mode has been designed to perform write or read operation under accelerated stress conditions (i.e., voltages higher

than 1.2V). For this, the bitlines and the wordline of a given cell are connected to their respective bitline and wordline analog paths, while the supply voltage terminal is connected to the stress supply voltage path.

- Standby mode: this operation mode has been designed to keep all cell terminals at ground voltage, therefore avoiding any aging degradation. For this, the bitline, wordline and power supply terminals are connected to their respective standby paths. This mode will be used when the cells are neither being measured, nor stressed.

Each unit cell in the array contains a total of 13 transmission gates that determine to which analog paths each terminal is connected, and, therefore, the operation mode of the cell. These transmission gates are (see Fig. 5.3):

- Each bitline has three transmission gates: two of them to connect the bitline node to the Force and Sense bitline analog paths, and a third one to connect it to the standby path.
- The wordline has two transmission gates: one to connect the node to the analog wordline path (where no Force-&-Sense structure is necessary, as the path is connected to transistor gates and no voltage drop is expected), and another one to connect the node to the standby path.
- The cell power supply node has five transmission gates: two for the force and sense nominal power supply paths, two others for the force and sense stress power supply paths of the stress, and one more for the standby connection. Analogously to the Endurance chip, stress parallelization can be achieved by connecting a number of cells to the stress power supply path, while the cell that has to be measured can be operated at nominal conditions through the nominal supply voltage path. This feature will be very useful to reduce the duration of the stress tests, which will be used to evaluate the reliability of the PUF under circuit degradation.

These transmission gates are controlled by a set of four control bits: XBL, which controls the connection of the access transistors to the bitline paths or to the standby path, as well as the connection of the cell power supply to the standby path; XWL, which controls the connection of the gates of the access transistors to the wordline or the standby path; XFP, which controls the connection of the cell power supply terminal to the nominal power supply path, and XHP, which controls the connection of the cell power supply to the stress power supply path. Table 5.1 displays the values at which the control bits must be set in order to enable each of the possible operation modes.

Operation mode	*XBL*	*XWL*	*XFP*	*XHP*
Measure	1	1	1	0
Stress hold	1	0	0	1
Stress write/read	1	1	0	1
Standby	0	0	0	0

Table 5.1: Values at which the four control bits have to be set in order to enable each of the possible operation modes for a given cell in the array.

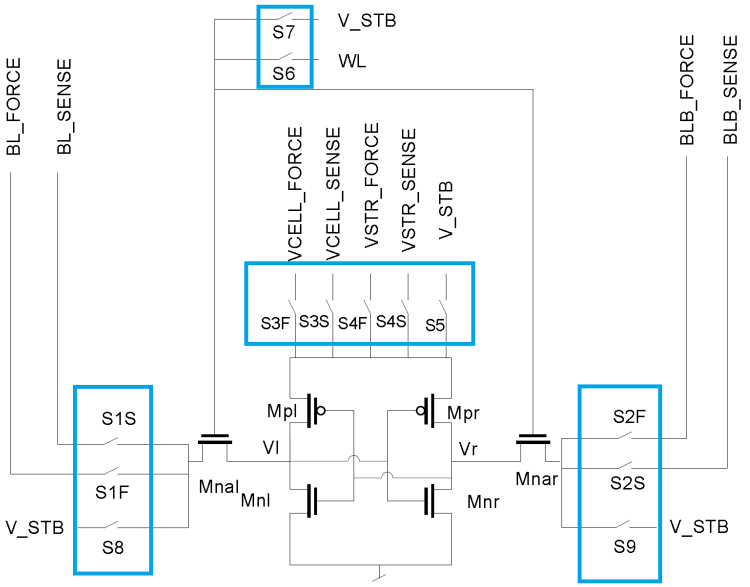

Figure 5.3: Schematic representation of the SRAM cell with the transmission gates for the connection of the analog paths to the cell terminals.

Fig. 5.4 displays a schematic representation of the SRAM unit cell in the KipT chip, showing the Circuit Under Test (CUT), the *ROW* and *COL* signals for the cell selection, and the four control bits used to control the transmission gates and determine to which analog paths the different SRAM cell terminals are connected. Notice that a set of level shifters has been included in the unit cell to provide the necessary voltage shift from the standard 1.2V supply voltage of the digital circuitry to the 3.3V operation voltage of the I/O transistors of the transmission gates. Fig. 5.5 shows the corresponding layout of the unit cell, with its main elements indicated.

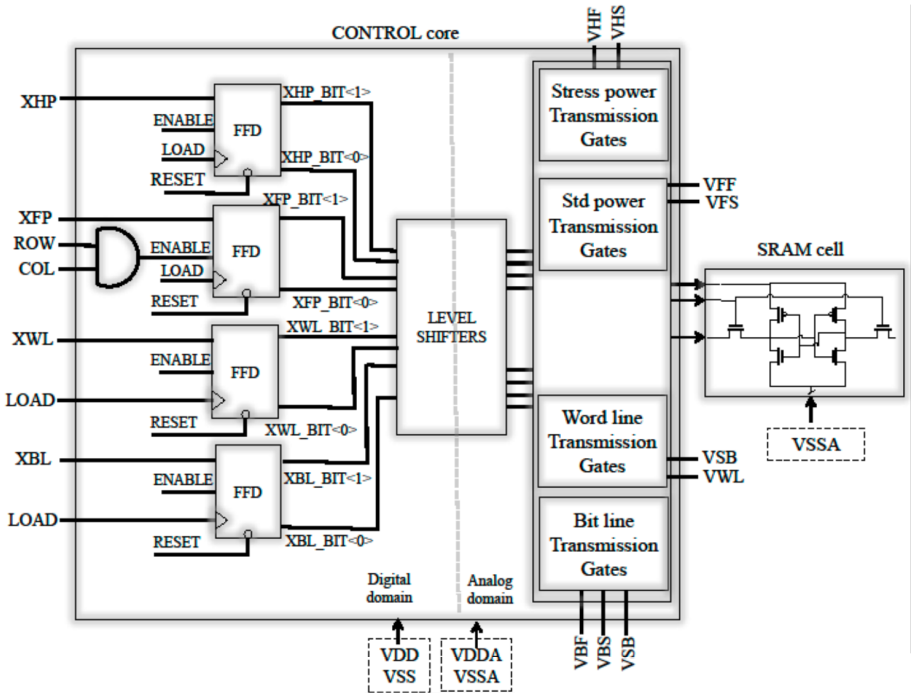

Figure 5.4: Block diagram of the SRAM unit cell in the KipT chip.

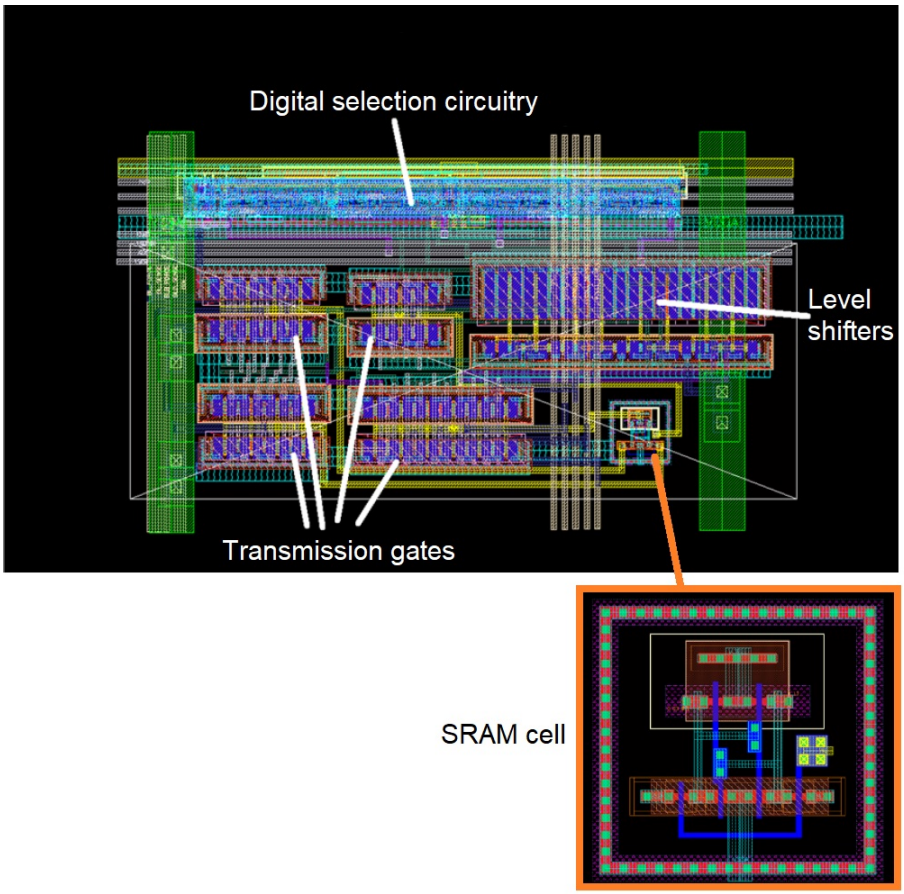

Figure 5.5: Layout of the SRAM unit cell in the KipT chip.

5.2 PUF properties and metrics

The main properties that a PUF should feature have been enumerated in Section 1.6.2. In this Section, a number of metrics to quantitatively evaluate those properties are provided. A broader description of the PUF properties and the corresponding metrics can be found in [129] and [130].

5.2.1 Bit Error Rate

The Bit Error Rate (BER) is a straightforward metric to assess the reliability of the PUF response. The BER can be defined as the fraction of erroneous bits (i.e., bits that differ from the reference or "golden" response) in a PUF response:

$$BER = \frac{n_{errors}}{N} \tag{5.1}$$

where n_{errors} denotes the number of erroneous bits, and N the total number of bits. It is clear that a perfectly reliable PUF that always returns the exact same response when the same challenge is applied to it has BER = 0.

5.2.2 Hamming Distance

The Hamming Distance (HD) is a common measure for the difference between two bit strings. If two binary vectors x and y of equal length N are considered, the Hamming Distance between them can be defined as the number of different bits that the two vectors have. This can be represented with an XOR operation as

$$HD(x, y) = \sum_{i=1}^{N} x_i \oplus y_i \tag{5.2}$$

Then, the normalized Hamming Distance between those two vectors can be simply defined as

$$HD_{norm}(x, y) = \frac{HD(x, y)}{N} \tag{5.3}$$

5.2.2.1 Inter Hamming Distance

The Inter Hamming Distance (Inter HD) corresponds to the Hamming Distance between the responses generated by different PUF instances to the same challenge and is therefore a measure of the PUF uniqueness. In the ideal case, the normalized

Inter HD values should lie around 0.5, since the individual bits generated by two different PUF instances (e.g., the power-up values from two SRAM cells in two different chips) should have a 50% chance of being equal.

If k different PUF instances are considered, their mean Inter HD can be defined as the mean of the Hamming Distance between the golden responses of each possible pair of instances [131]:

$$mean_{InterHD} = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{HD(R_i, R_j)}{N} \tag{5.4}$$

where N represents again the number of bits in each response and R_l represents the golden response of the l-th PUF instance.

5.2.2.2 Intra Hamming Distance

The Intra Hamming Distance (Intra HD) is defined as the Hamming Distance between two responses generated by the same PUF instance to a given challenge in different time instants and is therefore a metric for the PUF reliability. Ideally, a very reliable PUF instance would always return the same response to a given challenge, even when factors such as temperature variations or circuit degradation are taken into accout. Such an ideal PUF would present $IntraHD = 0$ when any two responses to the same challenge are considered.

Additionally, it is possible to define the maximum Intra HD as the maximum Hamming Distance between any pair of responses of the PUF instance to the same challenge, which would correspond to a worst-case scenario:

$$max_{IntraHD} = max\left(\frac{HD(R_i, R_j)}{N}\right) \text{ for all responses i, j} \tag{5.5}$$

5.2.3 Percentage of (un)stable bits

A stable bit within a PUF response can be defined as a bit (e.g., one of the SRAM cells in a SRAM array from which the PUF is built upon) that *always* yields the same value. In contrast, an unstable bit can be defined as a bit that does not always display the same value, i.e., there are at least two responses in which this bit has a different value.

$$Stable\ bits(\%) = \frac{n_{stable}}{N} \cdot 100 \tag{5.6}$$

$$Unstable\ bits(\%) = \frac{n_{unstable}}{N} \cdot 100 \tag{5.7}$$

In the case of SRAM power-up PUFs, stable bits or cells would correspond to cells that always power up to the same state, while unstable bits or cells would correspond that cells that do not always power up to the same value.

5.2.4 Hamming Weight

The Hamming Weight (HW) of a binary vector can be defined as the number of ones in it. It can be normalized just by dividing it by the total number of bits, as follows:

$$HW_{norm} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5.8}$$

where N represents the total number of bits, and x_i each of the bits.

5.2.5 Minimum Entropy

The Minimum Entropy (H_{min}) evaluates the randomness of the response of a PUF, and therefore its adequacy to generate True Random Numbers. It can be precisely defined as a lower bound on the entropy of the response, i.e., a worst-case measure of its unpredictability [132].

Consider an SRAM cell with a probability p_0 to power up to a '0', and a probability p_1 to power up to '1'. Let us define p_{max} as the greater of these two probabilities, i.e., $p_{max} = \max(p_0, p_1)$. Then, H_{min} of this cell can be calculated as

$$H_{min\_cell} = -log_2(p_{max}) \tag{5.9}$$

Consider the two extreme cases in terms of randomness for an SRAM cell. In the case of no randomness at all, the cell is stable and always powers up to a given value, e.g., to '0', then $p_{max} = p_0 = 1$, and therefore $H_{min} = 0$ for this cell. On the other hand, in the case of maximum randomness, the cell powers up to '0' half of the times and to '1' the other half, therefore $p_{max} = p_0 = p_1 = 0.5$, and therefore $H_{min} = 1$. From these lower and upper bounds, it becomes clear that a high H_{min} is desirable for True Random Number Generation.

If instead of a single cell, a string of N cells is considered, and it is assumed that their power-up values are independent, the minimum entropy of this sequence can be calculated as

$$H_{min} = -\frac{1}{N}\sum_{i=1}^{N} log_2(p_{i\_max}) \tag{5.10}$$

5.3 Previous Dark-Bit Masking methods to improve the reliability of SRAM PUFs

In this Section, some existing Dark-Bit Masking methods that aim at improving the reliability of SRAM PUFs are reviewed.

5.3.1 Multiple Evaluation approach

The Multiple Evaluation (ME) approach is maybe the most straightforward method for Dark-Bit Masking. The basic idea behind it is to perform several power-up evaluations to the SRAM cells of the PUF to classify them according to the strength of their power-up response [77], so that only the cells that always power up to the same value are used for key generation or entity authentication. An important feature of this method is that a finite number of evaluations must be selected. For instance, in [77], 20 power-up evaluations per cell was chosen as a good trade-off. However, the authors in [77] also showed that many cells that returned the same value during the first 20 power-ups (and hence would have been classified as completely stable in terms of their power-up response) returned at least one "erroneous" power-up value (i.e., their non-preferred power-up value) during the remaining 60 evaluations, since 80 evaluations per cell were performed in total in that work. Fig. 5.6 shows the results obtained for a ME classification using an SRAM array from the KipT chip [128]. The outcome of this test shows that using a number of evaluations in the order proposed in [77] would lead to labelling some cells as "stable", even if they present a relatively high probability of powering up to their non-preferred value.

There is another limitation to the ME method. As will be seen in the experimental data presented in 5.7, some cells that display a perfectly stable power-up behavior when fresh or pristine, and are therefore classified as stable by the ME procedure, may lose their stability after circuit aging.

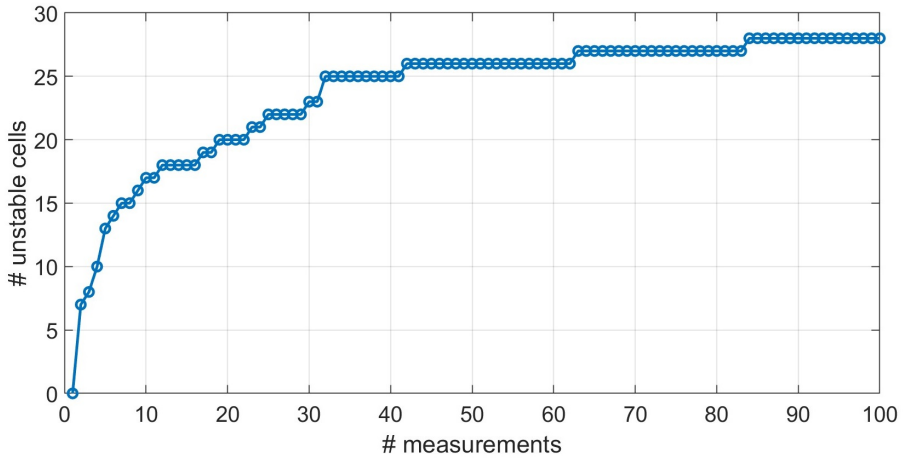

Figure 5.6: Number of unstable cells (i.e., cells that have at least one erroneous power-up value) against the number of power-up evaluations.

Apart from selecting the cells that always display the same power-up value for entity authentication or key generation, the ME procedure also allows to select the cells that display an unstable behavior (i.e., powering up to different values in different evaluations) for True Random Number Generation.

5.3.2 Data Remanence approach

Another interesting approach to increase the reliability of SRAM PUFs by Dark-Bit Masking was presented in [133]. The technique presented in that work consists of two remanence tests.

First, a '0' is written in each SRAM cell in the array, and the supply voltage is powered-off for a "short" time before being powered up again. Once the SRAM has been powered up again, the content of each cell is recorded. This is repeated several times changing the duration of the power-off interval. Then, the strength of the tendency of each cell towards '1' can be measured through the power-off time that it needs before it flips its content from '0' to '1', i.e., cells with a very strong tendency towards '1' will need a shorter power-off time to flip their content to '1', while cells with a weaker tendency towards '1' will need a longer power-off time to flip, or will not flip at all.

Then, the above-described procedure is repeated, this time writing a '1' in all cells in order to measure the strength of their tendency towards '0'.

This method has proven to be effective for experiments performed under different temperatures and power-up times, as well as under circuit aging [133]. This validation has been performed using a design fabricated in an ultra-low leakage technology, which requires relatively long power-down times (in the order of hundreds of milliseconds) to observe the cell bit flips. However, this approach may present some limitations in SRAMs designed in advanced CMOS technologies, where much shorter remancence times (e.g., below microseconds) are expected [133], [134], something that may turn this technique impractical in some technologies. Additionally, if power-on and power-off intervals in the order of microseconds, or even lower, need to be precisely controlled, the ramp rates must be much faster than that. This may pose a problem, since power-up states are very sensitive to factors such as noise when the ramp rates are in the range of nanoseconds to microseconds [99]. A possibility to overcome this issue would be to operate the Dark Bit Masking procedure at low temperatures, where much longer remanence times are expected [134], [135]. However, in order to significantly increase these remanence times, temperatures in the order of tens of degrees Celsius below zero would be needed, which usually involve the utilization of liquid nitrogen as a cooling method. This would make the technique very unpractical in real applications.

5.3.3 Exploiting the power supply ramp rate

Another method for the improvement of SRAM PUFs through Dark-Bit Masking has been presented in [99]. The main idea in this work revolves around the dependence that the power-up state of an SRAM cells has on the characteristics of the four core transistors. In this context, three different scenarios are considered. First, if the SRAM cell is powered on by ramping its V_{DD} node from low to high at a rapid rate, the power-up state of the cell will be determined entirely by the threshold voltage mismatch between its PMOS pull-up transistors. On the other hand, if the cell is powered on by rapidly ramping its V_{SS} node from high to low, the power-up state of the cell will be entirely determined by the threshold voltage mismatch between its NMOS pull-down transistors. The third possible scenario is that in which a "slow" ramp is used. In it, both PMOS and NMOS pairs have an impact in the resulting power-up state.

From this theoretical starting point, the authors of [99] propose the following method to classify the cells according to their power-up strength: first, a fast-ramp is applied from low to high to the V_{DD} node of the SRAM to evaluate which is the preferred power-up state of each cell if only its PMOS transistors are taken into account. Second, an analogous fast-ramp is applied from high to low to the V_{SS} node of the SRAM to evaluate which would be the preferred power-up state of each cell if only its core NMOS transistors are considered. Considering the two possible power-up preferences of each cell as in ('determined-by-PMOS', 'determined-by-NMOS'), there are four possible combinations: ('0','0'), ('0','1'), ('1','0'), ('1','1'). These four possibilities are in theory equally probable. Therefore, approximately 50% of the cells are expected to have the same preferred power-up value, both when only their PMOS or their NMOS core transistors are considered. These cells are then selected for the next step of the procedure. The other ones are discarded as not having a very strong power-up tendency.

In the next step of this procedure, the power-up strength of the cells that have not been discarded is quantitatively evaluated. To this end, all the cells in the SRAM have a '0' ('1') written on them. Then, V_{DD} is slowly ramped down to a low voltage (e.g., 0.2V). Then, V_{DD} is taken back to its nominal value, and the content of the cell is read to check if a bit flip has occurred. This step is repeated several times while varying the value of the low voltage (e.g., 0.2V, 0.15V, 0.1V, etc). Then, the strength of each cell towards '1' ('0') is evaluated by considering how low V_{DD} must be taken before a bit flip occurs, i.e., cells with a strong tendency towards '1' will flip their content at a not-so-low V_{DD} value (e.g., 0.3V), while cells with a weaker tendency towards '1' will only flip it at a lower V_{DD} (e.g., 0.1V), or will not flip at all.

However, this approach presents some limitations. First, unlike other Dark-Bit Masking methods such as the Multiple Evaluation one, or the one based on Data Remanence, it requires the relization of ramps in two different nodes of the cells (V_{DD} and V_{SS}). Second, because of its exploitation of "slow" and "fast" ramps, it requires the implementation of ramp rates that are orders of magnitude apart. This translates into additional overhead if the method is to be implemented on-

chip. Furthermore, the utilization of fast ramps at both supply nodes may be counterproductive, since the fast ramps performed in [99] are in the order of 1ns, which results in a power-up state that is very sensitive to factors such as noise or device degradation. Additionally, some potentially very stable cells are discarded for the last step of the procedure. Finally, it is not confirmed if the method is adequate to select cells that are stable not only under nominal conditions, but also when factors such as temperature variations or circuit aging are considered, since only simulation results using a Predictive Technology Model (PMT) for a 32-nm bulk technology, with estimated variability parameters, are provided in [99], and these simulations do not include the above-mentioned factors, which can be critical to the reliability of the PUF.

5.4 Maximum Trip Supply Voltage method

The MTSV method proposed in this Thesis is based on the Data Retention Voltage (DRV) metric of SRAM cells. The DRV of an SRAM cell can be defined as the minimum supply voltage at which that cell retains its stored value. The MTSV method evaluates the strength of the power-up tendency of a given cell A as follows: first, its non-preferred power-up value is written on it. Then, its supply voltage V_{DD} is lowered to a value $V_{DD\_low\_A}$ for a given period of time, before raising it again to its nominal value and performing a read operation to check if the cell has flipped its content. This step is repeated several times while varying $V_{DD\_low\_A}$. Then, the strength of the tendency of A towards its preferred power-up value is assessed by the $V_{DD\_low\_A}$ value at which the cell flips its content from its non-preferred value to its preferred value. For the sake of illustration, consider two cells, A and B, that have '0' as their preferred power-up value. Then, to evaluate which of them has a stronger tendency towards '0', a '1' is written on both of them, and the procedure described above is applied, thereby finding that A flips its content to '0' at $V_{DD\_low\_A}$, while B does it at $V_{DD\_low\_B}$. Then, if $V_{DD\_low\_A} > V_{DD\_low\_B}$, A is said to have a stronger tendency towards '0' than B, i.e., stronger cells have a higher DRV when their non-preferred value is written on them.

Ideally, if a cell already flips its content when its supply voltage is lowered to a given V_{DD_low}, it will also flip its content when its supply voltage is lowered to any voltage below V_{DD_low}. However, this is not always the case. During the MTSV procedure, it is observed that some cells that flip their content at a given V_{DD_low} may not always flip their content at lower supply voltages. This occurs because those cells do not have a strong power-up tendency and are thus discarded as "random" during the MTSV procedure.

5.5 Experimental strategy

In order to test the adequacy of the MTSV method to classify the strength of the cells' power-up tendency, the SRAM cell array within the KipT chip has been utilized

[128]. Then, four different groups of 128 cells have been considered among the 832 cells in one of the samples:

- *First*: this group corresponds to a selection of cells that does not follow any specific criterion to improve the reliability of the PUF. In particular, it corresponds to the first 128 cells within the array.
- *Random*: this group does not follow a criterion to improve the reliability of the PUF either. Instead of that, it is formed by 128 cells randomly selected from the array.
- *ME*: this group is formed by the 128 cells deemed strongest by the ME method when 20 power-ups are performed. The ME classification is performed only once at nominal bias and temperature conditions. Notice that this method only discards a small fraction of the cells as unstable (see Fig. 5.6), and the remaining cells are classified as equally stable by it. Therefore, assuming that the ME method has classified more than 128 cells as stable (which has always been the case in the tests performed for this Thesis), a random selection of 128 cells must still be performed among those cells.
- *MTSV*: this group comprises the 128 classified as strongest by the MTSV selection method. The MTSV classification is performed only once at nominal bias and temperature conditions.

Additionally, *All* will denote the group corresponding to all the 832 cells in the chip. Notice that, while both the ME and the MTSV selections are performed only once and at nominal voltage and temperature conditions, the power-up response of the whole array is evaluated at different conditions. Then, the reliability of each of the four different selections is compared through their BER. The different conditions under which the chip is studied are:

- Nominal conditions: 2,000 power-ups have been performed at nominal voltage (i.e., 1.2V) and temperature (i.e., 25ºC) conditions.
- Voltage variations: 200 power-ups have been performed by taking V_{DD} to different values ranging $\pm$ 10% of its nominal value, i.e., from 1.08V to 1.32V, while keeping the temperature at 25ºC.
- Temperature variations: 200 power-ups have been performed at different temperatures, ranging from -20ºC to 40ºC, always setting V_{DD} at its nominal value.
- After accelerated aging: 200 power-ups have been performed at different instants in time after the application of accelerated aging to the circuit. More details about the applied stress and the subsequent circuit degradation will be provided in Subsection 5.7.4.

5.6 Some preliminary considerations

The goal of the work presented in this Chapter is not to introduce a complete and novel PUF implementation, but rather to examine a novel technique that aims at improving the reliability of an already existing family of PUFs, the SRAM power-up PUFs. However, since the tests used to this end use a real silicon prototype of an SRAM cell array, some preliminary verification analysis has been performed to ensure the adequacy of the prototype for this purpose. For instance, imagine that, when different instances of the chip are powered up, they would display the same PUF response, i.e., the cells in equal position in all the arrays displayed the same power-up value. This would indicate that the power-up preference of the cells does not follow a random pattern caused by TZV and would dispute the adequacy of the prototype to be used to investigate such PUF Dark-Bit Masking methods.

These preliminary tests have consisted in powering up the SRAM array of 5 different chip instances (#1, #2, #3, #4 and #5) 200 times at nominal conditions. Then, the Hamming Weight of their golden response has been computed for each of them (see Table 5.2), and the mean Inter HD between their golden responses has been calculated to be

$$mean_{InterHD} = 0.4980 \tag{5.11}$$

The results obtained for the HW across the different samples indicate that the power-up preference of the cells is well balanced, i.e., there are approximately the same cells that prefer powering up to '0' than to '1'. In the case of the Inter HD, the obtained result approaches the ideal case of 0.5, indicating that the power-up response of each chip is indeed unique, and that there is no correlation between the power-up response of different arrays.

Chip instance	#1	#2	#3	#4	#5	mean
HW	0.5157	0.5308	0.5342	0.4902	0.5083	0.5158

Table 5.2: Hamming Weight calculated for the power-up response of five different chip instances, and their mean value.

5.7 PUF response reliability under different conditions

5.7.1 Reliability under nominal conditions

In this test, 2,000 power-ups have been performed with one of the array samples, and the power-up value of each of the cells in the array has been recorded each time. The

20 first power-ups have been used to classify the cells according to the ME method and select the 128 cells that form the *ME* group. Then, a MTSV classification has been performed to form the *MTSV* group. Then, the reliability for the different groups (*First*, *Random*, *ME* and *MTSV*) has been calculated through their BER. The results are displayed in Table 5.3, together with the BER obtained when all the 832 cells within the array are considered.

Group	**BER**
All	$5.221 \cdot 10^{-3}$
First	$4.625 \cdot 10^{-3}$
Random	$6.398 \cdot 10^{-3}$
ME	$2.27 \cdot 10^{-4}$
MTSV	$1.2 \cdot 10^{-5}$

Table 5.3: BER obtained for all the cells of the array, and for the different groups considered, after 2,000 power-ups at nominal conditions

It can be seen how the groups of 128 cells *First* and *Random* display a BER comparable to the one obtained when all cells are considered. This can be expected, since no criterion for the improvement of reliability has been followed to compose these groups. In contrast, both the *ME* and *MTSV* groups show an appreciable improvement with respect to *All*. In the case of *ME* this improvement is of roughly 1 order of magnitude, while in the case of *MTSV* the improvement is even larger, and reaches 2 orders of magnitude. This already shows some of the limitations of the ME selection method. Notice that 20 power-ups have been used to evaluate the strength of the cells, which was deemed adequate in [77]. However, as can be seen in Fig. 5.6, there are cells that always display the same power-up value during the first 20 evaluations, and therefore would be classified as perfectly stable by the ME method but may show errors at later evaluations.

5.7.2 Reliability under supply voltage variations

To investigate the reliability of the PUF response under supply voltage variations, which could appear during real-life operation, 200 power-ups have been performed to the complete array by taking its supply voltage to values in a range of ± 10% the nominal voltage, i.e., from 1.08V to 1.32V. The temperature has been kept constant at 25^{o}C during all the tests. The BER obtained for the different groups is displayed in Table 5.4.

Notice that, when all the cells are considered, the overall BER of the array stays, for the whole range of supply voltages, very similar to the one obtained at nominal voltage. Such a small variation can be explained by the fact that, when the supply voltage is ramped up during the power-up, the final state of each cell is "decided" already at an early stage of the power-up, i.e., at a low value of the supply voltage.

Therefore, slightly varying the final value of the supply voltage does not have a significant impact on the power-up state. In any case, it can be seen in Table 5.4 how the MTSV method always brings an improvement in the BER of at least one order of magnitude in comparison to the ME selection.

V_{DD} (V)	*All*	*First*	*Random*	*ME*	*MTSV*
1.08	$5.355 \cdot 10^{-3}$	$3.164 \cdot 10^{-3}$	$9.492 \cdot 10^{-3}$	$1.172 \cdot 10^{-3}$	$7.8 \cdot 10^{-5}$
1.14	$5.788 \cdot 10^{-3}$	$4.023 \cdot 10^{-3}$	$1.0469 \cdot 10^{-2}$	$1.406 \cdot 10^{-3}$	0
1.26	$6.444 \cdot 10^{-3}$	$3.555 \cdot 10^{-3}$	$1.0117 \cdot 10^{-2}$	$4.30 \cdot 10^{-4}$	0
1.32	$7.148 \cdot 10^{-3}$	$3.750 \cdot 10^{-3}$	$1.0508 \cdot 10^{-2}$	$3.91 \cdot 10^{-4}$	$3.9 \cdot 10^{-5}$

Table 5.4: BER obtained for all the cells of the array, and for the different groups considered, after 200 power-ups taking V_{DD} to different values around its nominal one.

5.7.3 Reliability under temperature variations

Another source of reliability issues in real operation can be temperature variations. To investigate their impact on the power-up response, 200 power-ups have been performed to a complete array under temperatures ranging from -20ºC to 40ºC, always taking the supply voltage to its nominal value. The results obtained for the complete array, as well as for each of the considered groups, is displayed in Table 5.5.

When all cells are considered, it can be seen that temperature variations introduce many more errors (i.e., increase the BER) than the above-discussed supply voltage variations. In particular, the overall BER increases both when the temperature increases and decreases from the nominal temperature of 25ºC. In all cases, both the ME and the MTSV methods bring a significant improvement to the BER of the corresponding groups. In particular, the MTSV-selected group has a BER of 0 for all cases except for T = -20ºC.

T (ºC)	*All*	*First*	*Random*	*ME*	*MTSV*
-20	$2.4272 \cdot 10^{-2}$	$2.2578 \cdot 10^{-2}$	$1.6641 \cdot 10^{-2}$	$1.17 \cdot 10^{-4}$	$3.9 \cdot 10^{-4}$
0	$1.3267 \cdot 10^{-2}$	$8.125 \cdot 10^{-3}$	$9.844 \cdot 10^{-3}$	0	0
10	$9.344 \cdot 10^{-3}$	$5.234 \cdot 10^{-3}$	$9.844 \cdot 10^{-3}$	$7.8 \cdot 10^{-5}$	0
20	$8.057 \cdot 10^{-3}$	$3.789 \cdot 10^{-3}$	$9.766 \cdot 10^{-3}$	$1.95 \cdot 10^{-4}$	0
40	$9.615 \cdot 10^{-3}$	$7.656 \cdot 10^{-3}$	$1.2734 \cdot 10^{-2}$	$3.711 \cdot 10^{-3}$	0

Table 5.5: BER obtained for all the cells of the array, and for the different groups considered, after 200 power-ups performed at different temperatures.

5.7.4 Reliability after circuit aging

Another factor that can threaten the reliability of the SRAM PUF response is the aging of the circuit that occurs during the circuit operation. The transistors within an SRAM cell can degrade in different manners, depending on the cell operation. However, it must be noted that the degradation of the cell transistors, and therefore of the circuit, does not necessarily bring a decrease in the cell's power-up reliability and an increase in the BER. In fact, it is known that storing a certain value (e.g., '0') in the cell strengthens the tendency of the cell to power up towards the opposite value (e.g., '1'). Therefore, storing the non-preferred value in a cell can be used to strengthen the cell's tendency to power up towards its preferred value, in what is known as *directed aging* [136], [126].

The aim of this work is to test the adequacy of the novel MTSV method to improve the reliability of the PUF as compared the scenarios in which no reliability-improvement method is used at all, and in which the ME method is used. Therefore, in order to have a fair comparison, the cells in the different groups must be subject to an equal amount of stress. Plus, this degradation should potentially bring a decrease in the cells' power-up strength. For this reason, the stress applied to each of the cells in the array has been equal, and in the opposite direction to the one usually known as *directed aging*. That is, the preferred power-up value of each cell has been stored in it, so that the cell's tendency towards that power-up value is weakened. This is illustrated in Fig. 5.7 for a cell that has a '1' as its preferred power-up value (i.e., it has a tendency to power-up with its node Vl to a high voltage value). In order to degrade such a cell's power-up strength, its preferred power-up value is written in the cell and held. This would cause the transistor Mpl to suffer NBTI and Mnr to suffer PBTI, as can be seen in Fig. 5.7. Therefore, the pull-up strength of the left node and the pull-down strength of the right node would be weakened, which also weakens the tendency of the left node to power-up to a high voltage value, and of the right node to power-up to a low voltage value (i.e., the tendency of the cell to power-up to a '1'). Following the previous explanation, the first step taken to degrade the cells' power-up tendency has been to write on each cell its preferred power-up value.

To achieve a noticeable degradation by storing the preferred power-up value at nominal conditions would require very long experimental times, in the range of weeks, months or even years. To speed up this process, accelerated aging has been applied by storing that value at a high supply voltage, in particular at $V_{DD} = 2.5V$ during 10,000s.

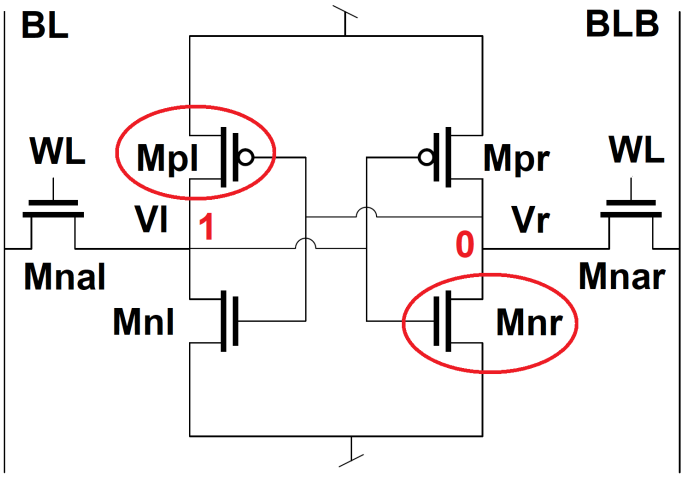

Figure 5.7: Transistors degraded during a hold stress with a '1' stored.

For the sake of illustration, let us consider the case of a cell that has '1' as its preferred power-up value. Then, the accelerated aging in this cell has been achieved by writing a '1', and storing it at V_{DD} = 2.5V during 10,000s. It must be noted that the BTI degradation undergone by the transistors in this case has both a permanent and a recoverable component, which starts to recuperate once the stress is removed. This recovered degradation follows a logarithmic behavior in the time scale [137]. To account for this recovery of the degradation, 200 power-ups have been performed at different times after the removal of the stress: right after the stress, 8 days after the stress, and 21 days after the stress. Notice that performing 200 power-ups and reading each power-up value takes a total of approximately 30s. Due to the aforementioned logarithmic temporal nature of BTI recovery, a considerable amount of degradation recuperates during the first 30s right after the removal of the stress, which means that the first measurement correspond to a non-stationary situation. On the other hand, the measurements performed 8 and 21 days after the stress correspond to a situation in which the degradation that recuperates during the 30s of measurement is negligible, and the situation can be considered stationary.

Note also that, if this stress were applied serially, the duration of the stress phase would be 832 cells $\times$ 10,000s/cell $\approx$ 96 days. However, it has already been said that the SRAM cell array of the KipT chip allows the application of stress in a parallel manner analogously to how the Endurance chip can do it. This is achieved by setting a number of cells in the Stress Hold operation mode by applying their power supply voltage through their stress path, while a single cell is kept in the Measure mode through the measurement path (see Section 5.1.1), in which it is possible to perform power-ups. Considering that only one cell can undergo power-ups at a time, and that it is important that all cells undergo exactly the same stress timing so that the results from different cells are equivalent, the parallelization scheme has been the following one. The first cell has been selected, its preferred value has been written on it, and the cell has been set in Stress hold mode with a supply voltage of 2.5V. Then, after 30s, which is the time that it takes to perform 200 power-ups, the second cell has been selected, its preferred value has been written on it, and the cell has

been taken to Stress hold mode. After this, the same procedure has been repeated consecutively for each cell in the array. When the first cell has been stressed for 10,000s, the first cell has been set to Measure mode and it has been powered up 200 times. After these 200 power-ups, which take 30s, the second cell undergoes the same procedure. This procedure is repeated for each cell consecutively. In this manner, all cells undergo the same stress time (10,000s) and are powered up exactly when their stress ceases. This procedure is depicted in a schematic manner in Fig. 5.8 for the first three cells of the array.

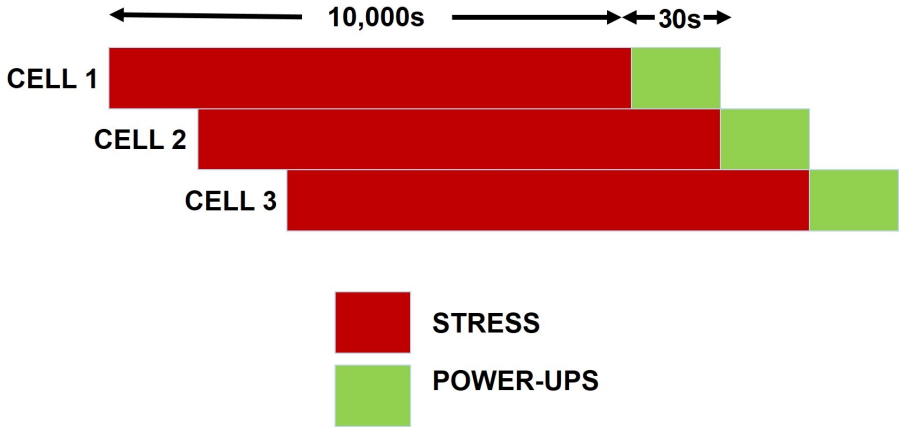

Figure 5.8: Schematic representation of the application of stress to the SRAM cells in a parallelized manner.

The results obtained for the BER of the whole array and for each of the groups, before the stress and at each of the tests performed after the application of stress, are displayed in Table 5.6. As expected, the overall BER of the array increases significantly after the application of the stress. Then, interestingly, the expected recovery of the BTI degradation at transistor level translates into a recovery (i.e., decrease) of BER with time. As occurred for nominal conditions, the *First* and *Random* groups display a similar evolution to the complete array, which is expected, since no selection method based on the improvement of reliability was used to make those groups. Although the BER obtained for the *ME* group is always better than the one obtained for those groups, it still degrades a lot after the application of the stress and, even if it recovers with time, it remains much higher than its pre-stress value. On the other hand, the *MTSV* group does not only show a better BER with respect to any other group right after the stress, but, once the recoverable component of the stress starts recuperating, and a static situation is achieved, it is at least 3 orders of magnitude better than any other selection, including the ME-based one. In particular, when 21 days have passed since the application of the stress, no power-up errors are recorded within the *MTSV* group, and its BER is 0.

Time	***All***	***First***	***Random***	***ME***	***MTSV***
Before	$7.858 \cdot 10^{-3}$	$6.836 \cdot 10^{-3}$	$1.445 \cdot 10^{-2}$	$4.3 \cdot 10^{-4}$	0
Right after	$7.405 \cdot 10^{-2}$	$6.949 \cdot 10^{-2}$	$6.430 \cdot 10^{-2}$	$4.652 \cdot 10^{-2}$	$1.07 \cdot 10^{-2}$
8 days after	$4.792 \cdot 10^{-2}$	$5.844 \cdot 10^{-2}$	$4.691 \cdot 10^{-2}$	$1.930 \cdot 10^{-2}$	$3.9 \cdot 10^{-5}$
21 days after	$4.714 \cdot 10^{-2}$	$5.141 \cdot 10^{-2}$	$4.672 \cdot 10^{-2}$	$1.223 \cdot 10^{-2}$	0

Table 5.6: BER obtained for all the cells of the array, and for the different groups considered, for 200 power-ups performed before the application of stress, and at different instants after it.

Chapter 6

Conclusions

For decades, CMOS technologies have been continuously scaled. This scaling allows to integrate a larger number of transistors into the same chip area, which in turn results in an increase in processor speed, a reduction of the power consumption, and an overall reduction in the manufacturing cost, among other advantages. However, these advancements have occurred at the expense of a much larger variability of the transistor parameters, both at time zero, just after the manufacturing process (Time-Zero Variability), and along time during the circuit operation (Time-Dependent Variability). This change in the variability is not only quantitative, but often also qualitative. This is the case for some Time-Dependent Variability phenomena, which could previously be modeled as deterministic processes, but have to be modeled as stochastic as transistors enter the nanometer range.

In this context, this Thesis revolves around the study of variability in CMOS technologies. This is done by following two very distinct approaches: first, Time-Dependent Variability phenomena that originate from the trapping/detrapping of charge carriers in/from defects in the transistors, such as RTN and BTI, are studied. This study includes the experimental characterization of the phenomena, and the subsequent construction of a model to describe them. This model could then be integrated, for example, in a Time-Dependent Variability simulation tool, which would allow the evaluation, and, eventually, mitigation of the negative effects of these phenomena. The second approach to CMOS variability in this Thesis consists in the exploitation of Time-Zero Variability for hardware security applications. This is done through the study of SRAM PUFs, which exploit the unpredictable variability that occur in that type of memory cells to create unique circuit fingerprints.

Regarding the study of Time-Dependent Variability, the work in this Thesis has first focused on the characterization of RTN, which consists of sudden and discrete threshold voltage shifts in transistors caused by trapping/detrapping of charge carriers in defects. These threshold voltage shifts cause, in turn, discrete shifts in the transistor current. A novel technique based on the Maximum Likelihood Estimation method to detect the distinct current levels, and, from them, extract the RTN transitions, has been presented. This technique has allowed to extract the statistical distribution of the current (or, analogously, threshold voltage) shifts associated to

RTN defects at different bias conditions. Furthermore, although the extraction of the defect time constants is not the goal of this part of the work, it is shown that there is no correlation between the time constants and the associated amplitude of each defect. After this, the Maximum Current Fluctuation metric, a new metric to characterize a current trace that displays RTN, has been presented. This metric can be described as the difference between the cumulative envelopes of the trace. In this sense, if a transistor has high RTN activity (and, thus, many and/or large current shifts), the difference between the current trace envelopes will be large, so that the Maximum Current Fluctuation of a transistor is an accurate reflection of its RTN activity. Then, a methodology based on this metric has been developed to extract some of the main parameters that characterize RTN, in particular those related to the amplitude associated to the RTN defects, and the number of active RTN defects along time. The main advantages of this methodology are that it does not require any complex processing, and that it is able to account for defects with a small associated amplitude. Indeed, this methodology reveals that techniques based on the extraction of each individual defect may fail to detect defects with a very small associated amplitude, especially when this amplitude lies close to or below the experimental background noise level. The study of RTN is concluded with an analysis of the impact that the biasing conditions experienced by the transistors prior to their measurement have on the measurement results. It is shown that, indeed, not only the biasing conditions during the measurement must be taken into account, but those prior to it, since they can have a considerable impact on the measurement outcome. This is a relevant topic which has not been addressed in the literature.

The study of Time-Dependent Variability phenomena is continued through the determination of the distribution of the defect time constants. This task can be tackled independently from the determination of the defect amplitudes since they have been shown to be uncorrelated. The determination of the distribution of the defect time constants is done through the utilization of the data from accelerated aging tests. These tests are more convenient for this task than the RTN tests for two reasons. First, the application of high voltages that is typically performed in the accelerated aging tests allows a broader exploration of the time constant map than the RTN measurements at nominal bias conditions. This is very important, since the time constant distribution is expected to span across a large number of decades in the time scale. Secondly, the application of stress can be performed in parallel to a high number of devices. This allows to perform longer stress periods to the devices without excessively increasing the overall test duration. These longer stress periods allow to investigate defects with longer capture times than those that can be attained if serial tests are considered. Furthermore, the algorithmic strategy developed to determine the time constant distribution has been thoroughly explained. This step-by-step explanation allows the reader to grasp the high complexity of the process, and to understand all the important details involved. Finally, the distributions obtained for the Time-Dependent Variability parameters have been used to develop tools for the simulation of Time-Dependent Variability phenomena, and for the emulation of the experimental characterization of these phenomena. This emulation could be used, for example, for the development of experimental data analysis tools, alleviating the user from the need of performing such tests in the lab, which can be expensive and time consuming.

Regarding the exploitation of Time-Zero Variability for security applications, a novel method for the improvement of the reliability of SRAM PUFs has been presented. This method, called the Maximum Trip Supply Voltage method, uses the Data Retention Voltage of each SRAM cell to predict how stable its power-up value is. In this sense, cells with a higher Data Retention Voltage value are expected to have a more stable power-up value (i.e., are expected to have a higher reproducibility in the value to which they power up when the supply voltage is applied to the cell). This allows to discard unstable cells that threaten the reliability of the SRAM PUF response. This method has been tested experimentally by using a novel chip, the KipT chip, which contains an array of SRAM cells. The KipT chip possesses some features that make it especially suited for this task, such as the ability to accurately apply the desired voltages to the SRAM cell terminals through a Force-&-Sense architecture, or the possibility of performing accelerated aging cells in a parallelized manner, among others. The tests that aim to validate the adequacy of the Maximum Trip Supply Voltage method have been performed under nominal operation conditions, and when factors such as supply voltage variations, temperature variations and circuit aging are considered. For the latter, accelerated aging has been applied to the circuit. These tests have been performed in an analogous manner to the BTI tests performed at the device level in the previous parts of this Thesis. The Maximum Trip Supply method has been shown to successfully predict which cells have a strongest power-up tendency under all the previously discussed conditions.

In conclusion, this Thesis has dealt with a wide variety of aspects of variability in CMOS technologies. Hopefully, the work presented hereby will be the foundation of interesting research in the future and will help the reader to gain a deeper understanding on topics such as defect-centric modeling of Time-Dependent Variability phenomena. It must be noted that much of the work presented in this Thesis has been, or is about to be, published in international journals and conferences.

Bibliography

[1] Ming Li. "Review of advanced CMOS technology for post-Moore era". In: *Science China Physics, Mechanics and Astronomy* 55.12 (2012), pp. 2316–2325.

[2] Robert H Dennard et al. "Design of ion-implanted MOSFET's with very small physical dimensions". In: *IEEE Journal of Solid-State Circuits* 9.5 (1974), pp. 256–268.

[3] Marcel JM Pelgrom, Aad CJ Duinmaijer, and Anton PG Welbers. "Matching properties of MOS transistors". In: *IEEE Journal of solid-state circuits* 24.5 (1989), pp. 1433–1439.

[4] Georges Gielen et al. "Emerging yield and reliability challenges in nanometer CMOS technologies". In: *Proceedings of the conference on Design, automation and test in Europe*. 2008, pp. 1322–1327.

[5] A Toro-Frias et al. "Lifetime calculation using a stochastic reliability simulator for analog ICs". In: *2018 15th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*. IEEE. 2018, pp. 1–9.

[6] Kwok K Hung et al. "Random telegraph noise of deep-submicrometer MOSFETs". In: *IEEE Electron Device Letters* 11.2 (1990), pp. 90–92.

[7] James H Stathis and Sufi Zafar. "The negative bias temperature instability in MOS devices: A review". In: *Microelectronics Reliability* 46.2-4 (2006), pp. 270–286.

[8] Alexander Acovic, Giuseppe La Rosa, and Yuan-Chen Sun. "A review of hot-carrier degradation mechanisms in MOSFETs". In: *Microelectronics Reliability* 36.7-8 (1996), pp. 845–869.

[9] *Strategic Research Agenda 2020 for Electronic Components Systems*. URL: `https://aeneas-office.org/wp-content/uploads/2020/07/ECS-SRA2020_L.pdf` (visited on 02/10/2021).

[10] J Martin-Martinez et al. "Probabilistic defect occupancy model for NBTI". In: *2011 International Reliability Physics Symposium*. IEEE. 2011, XT–4.

[11] David Atienza et al. "Reliability-aware design for nanometer-scale devices". In: *2008 Asia and South Pacific Design Automation Conference*. IEEE. 2008, pp. 549–554.

[12] Basel Halak, Mark Zwolinski, and M Syafiq Mispan. "Overview of PUF-based hardware security solutions for the Internet of Things". In: *2016 IEEE 59th International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE. 2016, pp. 1–4.

[13] William Shockley. "Problems related top-n junctions in silicon". In: *Czechoslovak Journal of Physics* 11.2 (1961), pp. 81–121.

[14] Kelin J Kuhn et al. "Process technology variation". In: *IEEE Transactions on Electron Devices* 58.8 (2011), pp. 2197–2208.

[15] Samar K Saha. "Modeling process variability in scaled CMOS technology". In: *IEEE Design & Test of Computers* 27.2 (2010), pp. 8–16.

[16] Sani R Nassif. "Process variability at the 65nm node and beyond". In: *2008 IEEE Custom Integrated Circuits Conference*. IEEE. 2008, pp. 1–8.

[17] Peter A Stolk, Frans P Widdershoven, and DBM Klaassen. "Modeling statistical dopant fluctuations in MOS transistors". In: *IEEE Transactions on Electron devices* 45.9 (1998), pp. 1960–1971.

[18] Samar K Saha. "Compact MOSFET modeling for process variability-aware VLSI circuit design". In: *IEEE Access* 2 (2014), pp. 104–115.

[19] Bing J Sheu et al. "BSIM: Berkeley short-channel IGFET model for MOS transistors". In: *IEEE Journal of Solid-State Circuits* 22.4 (1987), pp. 558–566.

[20] Fikru Adamu-Lema. "Scaling and Intrinsic Parameter Fluctuations in nanoCMOS Devices". PhD thesis. University of Glasgow, 2005.

[21] Tibor Grasser et al. "The paradigm shift in understanding the bias temperature instability: From reaction–diffusion to switching oxide traps". In: *IEEE Transactions on Electron Devices* 58.11 (2011), pp. 3652–3666.

[22] Maria Toledano-Luque et al. "Degradation of time dependent variability due to interface state generation". In: *2013 Symposium on VLSI Technology*. IEEE. 2013, T190–T191.

[23] Mulong Luo et al. "Impacts of random telegraph noise (RTN) on digital circuits". In: *IEEE Transactions on Electron Devices* 62.6 (2014), pp. 1725–1732.

[24] Xinyang Wang et al. "Random telegraph signal in CMOS image sensor pixels". In: *2006 International Electron Devices Meeting*. IEEE. 2006, pp. 1–4.

[25] Koichi Fukuda et al. "Random telegraph noise in flash memories-model and technology scaling". In: *2007 IEEE International Electron Devices Meeting*. IEEE. 2007, pp. 169–172.

[26] N Tega et al. "Increasing threshold voltage variation due to random telegraph noise in FETs as gate lengths scale to 20 nm". In: *2009 Symposium on VLSI Technology*. IEEE. 2009, pp. 50–51.

[27] Takashi Matsumoto, Kazutoshi Kobayashi, and Hidetoshi Onodera. "Impact of random telegraph noise on CMOS logic delay uncertainty under low voltage operation". In: *2012 International Electron Devices Meeting*. IEEE. 2012, pp. 25–6.

[28] AKM Mahfuzul Islam, Tatsuya Nakai, and Hidetoshi Onodera. "Statistical analysis and modeling of random telegraph noise based on gate delay measurement". In: *IEEE Transactions on Semiconductor Manufacturing* 30.3 (2017), pp. 216–226.

[29] K Takeuchi et al. "Single-charge-based modeling of transistor characteristics fluctuations based on statistical measurement of RTN amplitude". In: *2009 Symposium on VLSI Technology*. IEEE. 2009, pp. 54–55.

[30] Simeon Realov and Kenneth L Shepard. "Analysis of random telegraph noise in 45-nm CMOS using on-chip characterization system". In: *IEEE Transactions on Electron Devices* 60.5 (2013), pp. 1716–1722.

[31] Kenichi Abe et al. "Understanding of traps causing random telegraph noise based on experimentally extracted time constants and amplitude". In: *2011 International Reliability Physics Symposium*. IEEE. 2011, 4A–4.

[32] P Saraza-Canflanca et al. "A detailed study of the gate/drain voltage dependence of RTN in bulk pMOS transistors". In: *Microelectronic Engineering* 215 (2019), p. 111004.

[33] T Nagumo et al. "New analysis methods for comprehensive understanding of random telegraph noise". In: *2009 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2009, pp. 1–4.

[34] G Nicosia et al. "Investigation of the temperature dependence of random telegraph noise fluctuations in nanoscale polysilicon-channel 3-D Flash cells". In: *Solid-State Electronics* 151 (2019), pp. 18–22.

[35] James H Stathis, M Wang, and K Zhao. "Reliability of advanced high-k/metal-gate n-FET devices". In: *Microelectronics Reliability* 50.9-11 (2010), pp. 1199–1202.

[36] Sufi Zafar et al. "A comparative study of NBTI and PBTI (charge trapping) in SiO2/HfO2 stacks with FUSI, TiN, Re gates". In: *2006 Symposium on VLSI Technology*. IEEE. 2006, pp. 23–25.

[37] Kjell O Jeppson and Christer M Svensson. "Negative bias stress of MOS devices at high electric fields and degradation of MNOS devices". In: *Journal of Applied Physics* 48.5 (1977), pp. 2004–2014.

[38] Vincent Huard, M Denais, and C Parthasarathy. "NBTI degradation: From physical mechanisms to modelling". In: *Microelectronics Reliability* 46.1 (2006), pp. 1–23.

[39] Hans Reisinger et al. "Analysis of NBTI degradation-and recovery-behavior based on ultra fast VT-measurements". In: *2006 IEEE International Reliability Physics Symposium Proceedings*. IEEE. 2006, pp. 448–453.

[40] Tibor Grasser et al. "A two-stage model for negative bias temperature instability". In: *2009 IEEE International Reliability Physics Symposium*. IEEE. 2009, pp. 33–44.

[41] Tibor Grasser et al. "A unified perspective of RTN and BTI". In: *2014 IEEE International Reliability Physics Symposium*. IEEE. 2014, 4A–5.

[42] Ben Kaczer et al. "Atomistic approach to variability of bias-temperature instability in circuit simulations". In: *2011 IEEE International Reliability Physics Symposium*. IEEE. 2011, XT–3.

[43] James H Stathis, Souvik Mahapatra, and Tibor Grasser. "Controversial issues in negative bias temperature instability". In: *Microelectronics Reliability* 81 (2018), pp. 244–251.

[44] Ben Kaczer et al. "Ubiquitous relaxation in BTI stressing—New evaluation and insights". In: *2008 IEEE International Reliability Physics Symposium*. IEEE. 2008, pp. 20–27.

[45] Chenming Hu et al. "Hot-electron-induced MOSFET degradation-model, monitor, and improvement". In: *IEEE Journal of Solid-State Circuits* 20.1 (1985), pp. 295–305.

[46] Simon Tam, Ping-Keung Ko, and Chenming Hu. "Lucky-electron model of channel hot-electron injection in MOSFET's". In: *IEEE Transactions on Electron Devices* 31.9 (1984), pp. 1116–1125.

[47] Wenping Wang et al. "Compact modeling and simulation of circuit reliability for 65-nm CMOS technology". In: *IEEE Transactions on Device and Materials Reliability* 7.4 (2007), pp. 509–517.

[48] A Bayer et al. "Channel hot-carriers degradation in MOSFETs: A conductive AFM study at the nanoscale". In: *2013 IEEE International Reliability Physics Symposium (IRPS)*. IEEE. 2013, pp. 5D–4.

[49] Sufi Zafar et al. "A model for negative bias temperature instability (NBTI) in oxide and high/spl kappa/pFETs 13/spl times/-C6D8C7F5F2". In: *2004 Symposium on VLSI Technology, 2004.* IEEE. 2004, pp. 208–209.

[50] A Toro-Frias et al. "Reliability simulation for analog ICs: Goals, solutions, and challenges". In: *Integration, the VLSI Journal* 55 (2016), pp. 341–348.

[51] Paul Pfäffli et al. "TCAD for reliability". In: *Microelectronics Reliability* 52.9-10 (2012), pp. 1761–1768.

[52] Paul Pfäffli et al. "TCAD modeling for reliability". In: *Microelectronics Reliability* 88 (2018), pp. 1083–1089.

[53] V Velayudhan et al. "TCAD simulation of interface traps related variability in bulk decananometer mosfets". In: *2014 European Workshop on CMOS Variability (VARI)*. IEEE. 2014, pp. 1–6.

[54] Pieter Weckx et al. "Defect-centric perspective of combined BTI and RTN time-dependent variability". In: *2015 IEEE International Integrated Reliability Workshop (IIRW)*. IEEE. 2015, pp. 21–28.

[55] P Martin-Lloret et al. "CASE: A reliability simulation tool for analog ICs". In: *2017 International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*. IEEE. 2017, pp. 1–4.

[56] Takuya Komawaki et al. "Circuit-level simulation methodology for random telegraph noise by using verilog-AMS". In: *2017 IEEE International Conference on IC Design and Technology (ICICDT)*. IEEE. 2017, pp. 1–4.

[57] Engin Afacan et al. “A deterministic aging simulator and an analog circuit sizing tool robust to aging phenomena”. In: *2015 International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*. IEEE. 2015, pp. 1–4.

[58] A Toro-Frias et al. “Generation of Lifetime-Aware Pareto-Optimal Fronts Using a Stochastic Reliability Simulator”. In: *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE. 2019, pp. 78–83.

[59] Zhao Chuan Lee et al. “An 8T SRAM With On-Chip Dynamic Reliability Management and Two-Phase Write Operation in 28-nm FDSOI”. In: *IEEE Journal of Solid-State Circuits* 54.7 (2019), pp. 2091–2101.

[60] Luigi Atzori, Antonio Iera, and Giacomo Morabito. “The internet of things: A survey”. In: *Computer Networks* 54.15 (2010), pp. 2787–2805.

[61] Fang Hu, Dan Xie, and Shaowu Shen. “On the application of the internet of things in the field of medical and health care”. In: *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*. IEEE. 2013, pp. 2053–2058.

[62] Mussab Alaa et al. “A review of smart home applications based on Internet of Things”. In: *Journal of Network and Computer Applications* 97 (2017), pp. 48–65.

[63] Li Da Xu, Wu He, and Shancang Li. “Internet of things in industries: A survey”. In: *IEEE Transactions on Industrial Informatics* 10.4 (2014), pp. 2233–2243.

[64] Richard Kirk. “Cars of the future: the Internet of Things in the automotive industry”. In: *Network Security* 2015.9 (2015), pp. 16–18.

[65] Ravikanth Pappu et al. “Physical one-way functions”. In: *Science* 297.5589 (2002), pp. 2026–2030.

[66] Yansong Gao, Said F Al-Sarawi, and Derek Abbott. “Physical unclonable functions”. In: *Nature Electronics* 3.2 (2020), pp. 81–91.

[67] Kai-Hsin Chuang et al. “A physically unclonable function using soft oxide breakdown featuring 0% native BER and 51.8 fJ/bit in 40-nm CMOS”. In: *IEEE Journal of Solid-State Circuits* 54.10 (2019), pp. 2765–2776.

[68] Chip-Hong Chang, Yue Zheng, and Le Zhang. “A retrospective and a look forward: Fifteen years of physical unclonable function advancement”. In: *IEEE Circuits and Systems Magazine* 17.3 (2017), pp. 32–62.

[69] Roel Maes and Ingrid Verbauwhede. “Physically unclonable functions: A study on the state of the art and future research directions”. In: *Towards Hardware-Intrinsic Security*. Ed. by Ahmad-Reza Sadeghi and David Naccache. Springer, 2010, pp. 3–37.

[70] Charles Herder et al. “Physical unclonable functions and applications: A tutorial”. In: *Proceedings of the IEEE* 102.8 (2014), pp. 1126–1141.

[71] Srinivas Devadas et al. “Design and implementation of PUF-based" unclonable" RFID ICs for anti-counterfeiting and security applications”. In: *2008 IEEE International Conference on RFID*. IEEE. 2008, pp. 58–64.

[72] Sasa Mrdovic and Branislava Perunicic. "Kerckhoffs' principle for intrusion detection". In: *Networks 2008-The 13th International Telecommunications Network Strategy and Planning Symposium*. IEEE. 2008, pp. 1–8.

[73] Sergei Skorobogatov. "Flash memory 'bumping'attacks". In: *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer. 2010, pp. 158–172.

[74] Dilip SV Kumar et al. "An in-depth and black-box characterization of the effects of laser pulses on ATmega328P". In: *International Conference on Smart Card Research and Advanced Applications*. Springer. 2018, pp. 156–170.

[75] *White paper. The reliability of SRAM PUF*. URL: `https://www.intrinsic-id.com/wp-content/uploads/2017/08/White-Paper-The-reliability-of-SRAM-PUF.pdf` (visited on 01/21/2021).

[76] Daniel E Holcomb, Wayne P Burleson, and Kevin Fu. "Power-up SRAM state as an identifying fingerprint and source of true random numbers". In: *IEEE Transactions on Computers* 58.9 (2008), pp. 1198–1210.

[77] Iluminada Baturone, Miguel A Prada-Delgado, and Susana Eiroa. "Improved generation of identifiers, secret keys, and random numbers from SRAMs". In: *IEEE Transactions on Information Forensics and Security* 10.12 (2015), pp. 2653–2668.

[78] Ian Goldberg and David Wagner. "Randomness and the Netscape browser". In: *Dr Dobb's Journal-Software Tools for the Professional Programmer* 21.1 (1996), pp. 66–71.

[79] Andrew Rukhin et al. *A statistical test suite for random and pseudorandom number generators for cryptographic applications*. Tech. rep. National Institute of Standards and Technology, 2001.

[80] Jeroen Delvaux et al. "Helper data algorithms for PUF-based key generation: Overview and analysis". In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 34.6 (2014), pp. 889–902.

[81] Yevgeniy Dodis, Leonid Reyzin, and Adam Smith. "Fuzzy extractors: How to generate strong keys from biometrics and other noisy data". In: *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer. 2004, pp. 523–540.

[82] Ari Juels and Martin Wattenberg. "A fuzzy commitment scheme". In: *Proceedings of the 6th ACM conference on Computer and Communications Security*. 1999, pp. 28–36.

[83] Matthias Hiller, Ludwig Kürzinger, and Georg Sigl. "Review of error correction for PUFs and evaluation on state-of-the-art FPGAs". In: *Journal of Cryptographic Engineering* 10 (2020), pp. 229–247.

[84] James DR Buchanan et al. "'Fingerprinting'documents and packaging". In: *Nature* 436.7050 (2005), pp. 475–475.

[85] Ghaith Hammouri, Aykutlu Dana, and Berk Sunar. "CDs have fingerprints too". In: *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer. 2009, pp. 348–362.

[86] Keith Lofstrom, W Robert Daasch, and Donald Taylor. "IC identification circuit using device mismatch". In: *2000 IEEE International Solid-State Circuits Conference*. IEEE. 2000, pp. 372–373.

[87] Jae W Lee et al. "A technique to build a secret key in integrated circuits for identification and authentication applications". In: *2004 Symposium on VLSI Circuits*. IEEE. 2004, pp. 176–179.

[88] Blaise Gassend et al. "Silicon physical random functions". In: *Proceedings of the 9th ACM conference on Computer and Communications Security*. 2002, pp. 148–160.

[89] Blaise Laurent Patrick Gassend. "Physical random functions". PhD thesis. Massachusetts Institute of Technology, 2003.

[90] Jorge Guajardo et al. "FPGA intrinsic PUFs and their use for IP protection". In: *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer. 2007, pp. 63–80.

[91] Ying Su, Jeremy Holleman, and Brian Otis. "A 1.6 pJ/bit 96% stable chip-ID generating circuit using process variations". In: *2007 IEEE International Solid-State Circuits Conference*. IEEE. 2007, pp. 406–611.

[92] A Alheyasat et al. "Weak and Strong SRAM cells analysis in embedded memories for PUF applications". In: *2019 Conference on Design of Circuits and Integrated Systems (DCIS)*. IEEE. 2019, pp. 1–6.

[93] Elena Ioana Vatajelu, Giorgio Di Natale, and Paolo Prinetto. "Towards a highly reliable SRAM-based PUFs". In: *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE. 2016, pp. 273–276.

[94] Sudhir Satpathy et al. "13fJ/bit probing-resilient 250K PUF array with soft darkbit masking for 1.94% bit-error in 22nm tri-gate CMOS". In: *2014 European Solid State Circuits Conference (ESSCIRC)*. IEEE. 2014, pp. 239–242.

[95] Javier Diaz-Fortuny et al. "A versatile CMOS transistor array IC for the statistical characterization of time-zero variability, RTN, BTI, and HCI". In: *IEEE Journal of Solid-State Circuits* 54.2 (2018), pp. 476–488.

[96] Toshiharu Nagumo et al. "Statistical characterization of trap position, energy, amplitude and time constants by RTN measurement of multiple individual traps". In: *2010 International Electron Devices Meeting*. IEEE. 2010, pp. 28–3.

[97] Hans Reisinger et al. "Understanding and modeling AC BTI". In: *2011 International Reliability Physics Symposium*. IEEE. 2011, 6A–1.

[98] MJ Kirton and MJ Uren. "Noise in solid-state microstructures: A new perspective on individual defects, interface states and low-frequency (1/f) noise". In: *Advances in Physics* 38.4 (1989), pp. 367–468.

[99] Wendong Wang et al. "Exploiting power supply ramp rate for calibrating cell strength in SRAM PUFs". In: *2018 IEEE Latin-American Test Symposium (LATS)*. IEEE. 2018, pp. 1–6.

[100] Pablo Saraza-Canflanca et al. "A robust and automated methodology for the analysis of Time-Dependent Variability at transistor level". In: *Integration, the VLSI Journal* 72 (2020), pp. 13–20.

[101] Javier Diaz-Fortuny et al. "Flexible setup for the measurement of CMOS time-dependent variability with array-based integrated circuits". In: *IEEE Transactions on Instrumentation and Measurement* 69.3 (2019), pp. 853–864.

[102] Pablo Saraza-Canflanca et al. "TiDeVa: A Toolbox for the Automated and Robust Analysis of Time-Dependent Variability at Transistor Level". In: *2019 International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*. IEEE. 2019, pp. 197–200.

[103] Pablo Saraza-Canflanca et al. "New method for the automated massive characterization of Bias Temperature Instability in CMOS transistors". In: *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE. 2019, pp. 150–155.

[104] Javier Martin-Martinez et al. "New weighted time lag method for the analysis of random telegraph signals". In: *IEEE Electron Device Letters* 35.4 (2014), pp. 479–481.

[105] Javier Martin-Martinez et al. "Characterization of random telegraph noise and its impact on reliability of SRAM sense amplifiers". In: *2014 European Workshop on CMOS Variability (VARI)*. IEEE. 2014, pp. 1–6.

[106] Victor M van Santen et al. "Weighted time lag plot defect parameter extraction and GPU-based BTI modeling for BTI variability". In: *2018 IEEE International Reliability Physics Symposium (IRPS)*. IEEE. 2018, P–CR.

[107] Javier Diaz-Fortuny et al. "A smart noise-and RTN-removal method for parameter extraction of CMOS aging compact models". In: *Solid-State Electronics* 159 (2019), pp. 99–105.

[108] Jian-Xin Pan and Kai-Tai Fang. "Maximum likelihood estimation". In: *Growth Curve Models and Statistical Diagnostics*. Springer, 2002, pp. 77–158.

[109] In Jae Myung. "Tutorial on maximum likelihood estimation". In: *Journal of Mathematical Psychology* 47.1 (2003), pp. 90–100.

[110] Jeffrey C Lagarias et al. "Convergence properties of the Nelder–Mead simplex method in low dimensions". In: *SIAM Journal on Optimization* 9.1 (1998), pp. 112–147.

[111] Javier Diaz-Fortuny et al. "A model parameter extraction methodology including time-dependent variability for circuit reliability simulation". In: *2018 International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*. IEEE. 2018, pp. 53–56.

[112] Zexuan Zhang et al. "Investigation on the amplitude distribution of random telegraph noise (RTN) in nanoscale MOS devices". In: *2016 IEEE International Nanoelectronics Conference (INEC)*. IEEE. 2016, pp. 1–2.

[113] Zexuan Zhang et al. "New insights into the amplitude of random telegraph noise in nanoscale MOS devices". In: *2017 IEEE International Reliability Physics Symposium (IRPS)*. IEEE. 2017, pp. 3C–3.

[114] Christian Monzio Compagnoni et al. "Three-dimensional electrostatics-and atomistic doping-induced variability of RTN time constants in nanoscale MOS devices—Part II: Spectroscopic implications". In: *IEEE Transactions on Electron Devices* 59.9 (2012), pp. 2495–2500.

[115] Herbert A David and Haikady N Nagaraja. *Order statistics*. 3rd ed. John Wiley & Sons, 2003.

[116] Samuel Stanley Wilks. "Order statistics". In: *Bulletin of the American Mathematical Society* 54.1 (1948), pp. 6–50.

[117] James Kennedy and Russell Eberhart. "Particle swarm optimization". In: *Proceedings of ICNN'95-International Conference on Neural Networks*. Vol. 4. IEEE. 1995, pp. 1942–1948.

[118] Runsheng Wang et al. "Too Noisy at the Bottom?—Random telegraph noise (RTN) in advanced logic devices and circuits". In: *2018 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2018, pp. 17–2.

[119] Amy Whitcombe et al. "On-chip I–V variability and random telegraph noise characterization in 28 nm CMOS". In: *2016 European Solid-State Device Research Conference (ESSDERC)*. IEEE. 2016, pp. 248–251.

[120] Marko Simicic et al. "Advanced MOSFET variability and reliability characterization array". In: *2015 IEEE International Integrated Reliability Workshop (IIRW)*. IEEE. 2015, pp. 73–76.

[121] Salvatore Maria Amoroso et al. "RTN and BTI in nanoscale MOSFETs: A comprehensive statistical simulation study". In: *Solid-State Electronics* 84 (2013), pp. 120–126.

[122] Kyosuke Ito et al. "Modeling of random telegraph noise under circuit operation—Simulation and measurement of RTN-induced delay fluctuation". In: *2011 International Symposium on Quality Electronic Design*. IEEE. 2011, pp. 1–6.

[123] P Martin-Lloret et al. "An IC Array for the Statistical Characterization of Time-Dependent Variability of Basic Circuit Blocks". In: *2019 International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*. IEEE. 2019, pp. 241–244.

[124] P Martın-Lloret et al. "A size-adaptive time-step algorithm for accurate simulation of aging in analog ICs". In: *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. 2017, pp. 1–4.

[125] Umer Hassan and Muhammad Sabieh Anwar. "Reducing noise by repetition: introduction to signal averaging". In: *European Journal of Physics* 31.3 (2010), p. 453.

[126] Roel Maes and Vincent Van Der Leest. "Countering the effects of silicon aging on SRAM PUFs". In: *2014 IEEE International Symposium on Hardware-Oriented Security and Trust (HOST)*. IEEE. 2014, pp. 148–153.

[127] J Nuñez et al. "Experimental Characterization of Time-Dependent Variability in Ring Oscillators". In: *2019 International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*. IEEE. 2019, pp. 229–232.

[128] Pablo Saraza-Canflanca et al. "Design Considerations of an SRAM Array for the Statistical Validation of Time-Dependent Variability Models". In: *2018 International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*. IEEE. 2018, pp. 73–76.

[129] Roel Maes. *Physically unclonable functions: Constructions, properties and applications*. Springer Science & Business Media, 2013.

[130] Abhranil Maiti, Vikash Gunreddy, and Patrick Schaumont. "A systematic method to evaluate and compare the performance of physical unclonable functions". In: *Embedded systems design with FPGAs*. Ed. by Peter Athanas, Dionisios Pnevmatikatos, and Nicolas Sklavos. Springer, 2013, pp. 245–267.

[131] Abhranil Maiti et al. "A large scale characterization of RO-PUF". In: *2010 IEEE International Symposium on Hardware-Oriented Security and Trust (HOST)*. IEEE. 2010, pp. 94–99.

[132] Elaine B Barker and John Michael Kelsey. *Recommendation for random number generation using deterministic random bit generators (revised)*. US Department of Commerce, Technology Administration, National Institute of Standards and Technology, Computer Security Division, Information Technology Laboratory, 2007.

[133] Muqing Liu et al. "A data remanence based approach to generate 100% stable keys from an sram physical unclonable function". In: *2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. IEEE. 2017, pp. 1–6.

[134] Cagla Cakir, Mudit Bhargava, and Ken Mai. "6T SRAM and 3T DRAM data retention and remanence characterization in 65nm bulk CMOS". In: *IEEE 2012 Custom Integrated Circuits Conference*. IEEE. 2012, pp. 1–4.

[135] Nikolaos Athanasios Anagnostopoulos et al. "Low-temperature data remanence attacks against intrinsic SRAM PUFs". In: *2018 Euromicro Conference on Digital System Design (DSD)*. IEEE. 2018, pp. 581–585.

[136] Mudit Bhargava, Cagla Cakir, and Ken Mai. "Reliability enhancement of bistable PUFs in 65nm bulk CMOS". In: *2012 IEEE International Symposium on Hardware-Oriented Security and Trust*. IEEE. 2012, pp. 25–30.

[137] Sanjay Rangan, Neal Mielke, and Everett CC Yeh. "Universal recovery behavior of negative bias temperature instability [PMOSFETs]". In: *IEEE International Electron Devices Meeting*. IEEE. 2003, pp. 14–3.