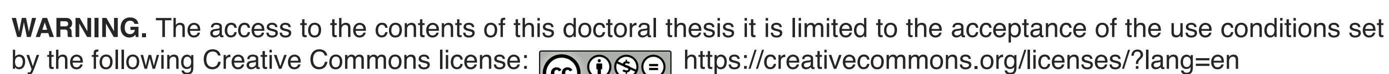
Universitat Autònoma de Barcelona

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons: http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons: http://es.creativecommons.org/blog/licencias/

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  https://creativecommons.org/licenses/?lang=en

Universitat Autònoma de Barcelona

A versatile framework for the statistical characterization of CMOS time-zero and time-dependent variability with array-based ICs

Author:
Javier Díaz Fortuny

Supervisors:
Dr. Javier Martín Martínez
Dr. Rosana Rodríguez Martínez

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Electrical and Telecommunication Engineering

in the

Reliability of Electron Devices and Circuits group (REDEC)
Department of Electronic Engineering

July 2019

Universitat Autònoma
de Barcelona

The undersigned, Dr. Javier Martín Martínez and Dr. Rosana Rodríguez Martínez, Associate Professors of the Electronic Engineering Department (Engineering School) of the *Universitat Autònoma de Barcelona*

CERTIFY:

that the thesis entitled "*A versatile framework for the statistical characterization of CMOS time-zero and time-dependent variability with array-based ICs*" has been written by the Ph.D candidate **Javier Díaz Fortuny** under their supervision, in fulfilment of the requirements for the Ph.D degree in Electrical and Telecommunication Engineering.

Author:	**Supervisors:**
Mr. Javier Díaz Fortuny	Dr. Javier Martín Martínez
	Dr. Rosana Rodríguez Martínez

Bellaterra (Barcelona), July 2019

You do not really understand something unless you can explain it to your grandmother.

Albert Einstein

Agradecimientos

Me gustaría comenzar mi tesis doctoral dedicando unas palabras de agradecimiento a todas aquellas personas que me han acompañado durante estos cuatro años de travesía.

En primer lugar, querría dar las gracias a mis directores de tesis, el Dr. Javier Martín Martínez y la Dra. Rosana Rodríguez Martínez, por guiarme y orientarme durante el largo camino recorrido durante mi tesis doctoral. Vuestro apoyo, ayuda e ideas compartidas a lo largo de estos cuatro años han sido imprescindibles para llegar a la culminación de esta tesis. También me gustaría agradecer la confianza depositada en mí a la Dra. Montserrat Nafría Maqueda, quien me ofreció la oportunidad de realizar mis estudios de doctorado dentro del grupo de investigación REDEC en el Departamento de Ingeniería Electrónica de la Universidad Autónoma de Barcelona.

Por otro lado, quisiera dar las gracias encarecidamente a mis compañeros de despacho, así como a todos los miembros del grupo REDEC. En especial quiero destacar a Alberto Rodríguez, Carlos Couso, Oscar Súchil, Miquel Moras, Cristian Herrojo y Marcos Maestro. Cada uno de vosotros, vuestra compañía y apoyo durante estos años, ha marcado un punto de inflexión imposible de olvidar. Nuestras discusiones, chascarrillos, bromas y excentricidades perdurarán en mi memoria para siempre. También quisiera extender mi gratitud a los miembros del Departamento de Ingeniería Electrónica por todos los buenos momentos que hemos compartido, en especial a Javier Hellín por su ayuda professional con el diseño gráfico de la tesis doctoral.

Una parte muy importante de esta tesis corresponde a la colaboración mantenida con el Instituto de Microelectrónica de Sevilla. En especial me gustaría destacar todo el apoyo recibido por parte de Rafael Castro López, Francisco V. Fernández y Elisenda Roca. Las innumerables horas invertidas en el laboratorio de caracterización y la generación de algoritmos y setups de medida, que culminaron con una curva I_{DS}–V_{GS}, junto con las intensas discu-

siones llevadas a cabo a lo largo de estos años, han contribuido enormemente a crear una relación profesional y personal difícil de superar. Por descontado, también me gustaría agradecer sinceramente a los que fueron mis compañeros durante mis estancias en Sevilla, en especial a Antonio Toro, Pablo Sarazá y Fabio Moreira. Nuestras discusiones e interminables skypes han contribuido notablemente al desarrollo de esta tesis doctoral.

También quisiera extender mis agradecimientos a Víctor Van Saten y Hussam Hamroch con quienes trabajé conjuntamente durante mi estancia de investigación en el Karlsruher Institute Für Technologie en Alemania. Una experiencia enriquecedora en la que dedicamos largas noches a la creación de algoritmos de procesado de datos e interesantes discusiones que han contribuido a mi crecimiento como investigador. Quisiera también transmitir mi mas sincero agradecimiento a todos los compañeros del grupo de investigación Device Reliability and Electrical Characterization group en IMEC, Bélgica. En especial a Dimitri Lintern, Erick Bury, Ben Kaczer, Marko Simicic, Md Nur Kutubul Alam y Geert Hellings. Vuestra acogida me permitió integrarme en vuestro grupo de investigación como si fuera uno más y adquirir nuevos y valiosos conocimientos. Asimismo me gustaría agradecer el apoyo y colaboración recibidos por el grupo de investigación HIPICS de la Universidad Politécnica de Catalunya durante la fase de diseño del chip.

Antes de terminar, quisiera dedicar todo el trabajo realizado a lo largo de mi vida académica a mi familia, en especial a mi madre Amalia Fortuny y a mi padre Francisco Javier Díaz por apoyarme y guiarme desde muy pequeño para continuar avanzando y ser competitivo. A Pedro Sánchez y a mi hermana Meritxell Díaz por apoyarme durante todos estos años de estudio. También quiero dedicar una mención especial a mis abuelos, Carme Bartolomé y Josep Fortuny, por impulsarme cada día a estudiar y creer siempre que lograría mis objetivos. Por último, quiero agradecer a mis suegros, Carmen Hinojosa y Herminio Albaladejo, su constante e inestimable apoyo y los ánimos que me han dado durante todos estos años. A todos ellos, gracias.

Finalmente, y desde lo más profundo de mi corazón, me gustaría otorgar mi más sincero agradecimiento a mi esposa, Gloria Albaladejo. De no haber estado a mi lado, con tu amor y apoyo incondicionales, nunca habría llegado tan lejos. No hay palabras para describir mi gratitud por todos estos años y por toda la vida que nos queda por vivir juntos. Te amo.

Javier Díaz Fortuny
Barcelona, Julio 2019.

Acknowledgements

I would like to start my PhD thesis extending a word of thanks to all the people who have supported me during these past four years.

Firstly, I would like to thank my thesis directors, Dr. Javier Martín Martínez and Dr. Rosana Rodríguez Martínez, for guiding me throughout my PhD thesis. Your support, help and ideas during these past four years have been indispensable for the completion of this thesis. I would also like to take this opportunity to recognize the trust placed in me by Dr. Montserrat Nafría Maqueda, who offered me the opportunity to carry out my PhD studies within the REDEC group in the Engineering Department in the Autonomous University of Barcelona.

Secondly, I would like to thank my fellow office colleagues, as well as to all REDEC group members. I would specially like to pay tribute to Alberto Rodríguez, Carlos Couso, Oscar Súchil, Miquel Moras, Cristian Herrojo and Marcos Maestro. Each one of you, your companionship and support, have set an inflexion point impossible to forget. Our discussions, jokes, and eccentricities will forever live in my memory. I would also like to extend my thanks to all Electronic Engineering Department members for all the good moments shared, especially to Javier Hellín, for his professional help with the graphic design of my PhD thesis.

A very important part of this thesis corresponds to the collaboration with the Microelectronic Institute of Seville. I would specially highlight the support I have been given by Rafael Castro López, Francisco V. Fernández and Elisenda Roca. The countless hours invested in the characterization laboratory and the algorithm and measurement setups generation, which led to a I_{DS}–V_{GS} curve, together with the intense discussions carried out during these years, have greatly contributed to create a professional and personal relationship difficult to top. It goes without saying that I sincerely thank my fellow colleagues during my stay at Seville, especially Antonio Toro,

Pablo Sarazá and Fabio Moreira. Our debates and endless skypes contributed significantly to the development of this PhD thesis.

I would also like to extend my gratitude to Víctor Van Saten and Hussam Hamroch, whom I worked with during my research stay at Karlsruher Institute Für Technologie in Germany. A rewarding experience in which we dedicated long nights to the creation of processing algorithms and maintained interesting debates, which have contributed to my growth as a researcher. I would also like to give my most sincere thanks to all the colleagues in the Device Reliability and Electrical Characterization group at IMEC, Belgium. Especially, to Dimitri Lintern, Erick Bury, Ben Kaczer, Marko Simicic, Md Nur Kutubul Alam and Geert Hellings. Your warm welcome allowed me to become part of your reasearch group, as well as to acquire new and valuable knowledge. Likewise, I would like to thank all the support and collaboration received by the HIPICS research group from the Politechnic University of Cataluña (UPC) during the design phase of the chip.

Before I finish, I would like to dedicate of the work done throughout my academic life to my family, especially to my mohter Amalia Fortuny and my father Francisco Javier Díaz, for supporting and guiding me since my early years in life to make further progress and to be competitive. To Pedro Sánchez and my sister Meritxell Díaz for supporting me throughout all these years of study. I would also like to make a special reference to my grandparents, Carme Bartolomé and Josep Fortuny, for pushing me every day to study and always believe that I would achieve my goals. Last but not least, I want to thank my in-laws, Carmen Hinojosa and Herminio Albaladejo, for their constant and invaluable support and for all their encouragement throughout the years. To all of them, thank you.

Finally, and from the bottom of my heart, I would like to express my sincerest thanks to my dear wife, Gloria Albaladejo. Had you not been by my side, with unconditional love and support, I would have never come this far. I cannot emphasize enough my gratitude for all these years and for those still to come. I love you.

Javier Díaz Fortuny
Barcelona, Julio 2019.

Preface

Since the invention in 1960 of the Metal Oxide Semiconductor Field Effect Transistor (MOSFET), the CMOS semiconductor industry has invariably invented new feats to progressively reduce the minimum gate length, from the micrometer scale ($\approx 10\mu m$) to the current 7-14-nm gate lengths or the new 5-nm technology node predicted to be manufactured in 2020-2021, all with the aim of fabricating more reliable devices and even more advanced circuits and systems, with billions of transistors per chip.

With all the benefits that transistor size scaling brings to power, area and performance, approaching the atomic scale poses, due to the very discreteness of matter, an important peril: the increase of variations of the transistor's intrinsic performance, thus critically compromising the fundamental reliability of the fabricated devices and circuits. In this way, variations of fabricated transistor parameters, like for instance threshold voltage or mobility, as well as their degradation during circuit functionality, have become an increasing concern in nanometer integrated circuit design. Moreover, a significant increase of gate leakage current has emerged due to the scaling in the thickness of the transistor's insulator. In this scenario, to increase performance and reliability of the fabricated devices, new and more complex stack materials have been introduced, such as silicon oxynitride (SiON), High-K Metal gate insulators (HKMG) and new device geometries like FinFETs, FDSOI or MuGFETs have emerged in ultra-scaled technology nodes to continue with the scaling trend and have better control of the short channel effects.

The variability in the transistor parameters, stochastic by nature, must be massively characterized in order to capture those variations with a representative and sound statistical sampling. Variability sources are divided in two different types: first, the time-zero variability, typically known as process variability which occurs during the fabrication process and consists in a permanent either random or systematic, shift of the device parameters; second, the time-dependent variability, which occurs during device or circuit

operation over time and includes transient effects like Random Telegraph Noise, and degradation mechanisms or aging effects, like Hot Carrier Injection, Bias Temperature Instability, Time Dependent Dielectric Breakdown, Stress Induced Leakage Current, etc., which are potential sources of device and IC variability that can lead transistors to a progressive degradation or to a permanent failure.

In order to reduce or mitigate variability effects, novel variability-aware circuit design techniques are required to assess the combined impact of time-zero and time-dependent variability in advanced technology nodes. Variability-aware techniques utilizes accurate compact models, which are based in statistical characterization of individual MOSFET devices. In this regard, providing statistically accurate characterization of time-zero and time-dependent variability effects in modern CMOS technologies has, therefore, become a key step in the path towards attaining truly reliable integrated circuits.

In this context, this thesis, structured in 4 chapters, will contribute to the characterization and lifetime prediction of nanometer CMOS technologies through a thorough study of extensive statistical data samples. To do so, issues related to typical serial characterization techniques, which require months or even years of continuous non-stop device testing, are overcome thanks to a novel and versatile array-based IC chip design in conjunction with a full-custom characterization framework. These two key elements, the IC and the framework, can effectively be utilized to statistically characterize the impact of different device variability sources in nanometer-scale MOSFETs while significantly and outstandingly reducing the required characterization time.

To gently introduce the reader to the material, in the first chapter of this thesis, a brief discussion of the most relevant time-zero and time-dependent variability phenomena in today's fabricated nano-MOSFETs is provided. The role that the interface states, present in the transistor insulator, plays in the device variability is also explained here. Moreover, the necessary requirements for both a capable array-based IC design and a trustworthy measurement framework for the statistical device characterization will be described. Finally, a state-of-the-art review of the pros and cons of the most recent array-based ICs reported in the literature, for the statistical characterization of variability phenomena, is presented highlighting the basic but important concept of parallel stress in IC chips.

In the context of getting enough statistical data to accurately characterize transistor variability while avoiding non-feasible, on-wafer serial device characterization times, the second chapter will describe the design of a novel array-based IC chip, named ENDURANCE, by which time-zero and time-dependent variability effects can be both observed and measured. To significantly cut down the time for aging phenomena characterization, the versatility of the ENDURANCE chip hardware, guaranteeing massively parallel aging testing, will be described.

The third chapter describes an innovative and automated measurement framework, named TARS, whose focus is to provide statistical measurements to help CMOS analog and digital circuit designers to fight against variability phenomena. Moreover, a novel and smart parallel aging testing technique that introduces a significant reduction of the on-wafer serial aging tests from years to hours will be described.

In chapter four, all data collected through the statistical measurements conducted using TARS sofware have been used to study the effects of time-zero and time-dependent variability phenomena towards attaining a precise prediction of the expected technology lifetime. Moreover, several novel methodologies utilized for the extraction of transistor model parameters intended for accurate circuit simulation will be also presented and described.

Finally, the most relevant conclusions of the work presented in this thesis are summarized.

Publications related to this thesis

International journal papers as first author:

1. 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." in IEEE Transactions on Instrumentation and Measurement, March 2019. DOI: 10.1109/TIM.2019.2906415 Impact factor (2017): 2.794. Quartile: Q1.
2. J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez, R. Castro-Lopez, E. Roca, X. Aragones, E. Barajas, D. Mateo, F. V. Fernandez and M. Nafria, "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, vol. 54, no. 2, pp. 476-488, February 2019. DOI: 10.1109/JSSC.2018.2881923. Impact factor (2017): 4.075. Quartile: Q1.
3. J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez, R. Castro-Lopez, E. Roca, F. V. Fernandez and M. Nafria, "A Smart Noise- and RTN-Removal Method for Parameter Extraction of CMOS Aging Compact Models" in Solid-State Electronics (Elsevier), March 2019. DOI: 10.1016/j.sse.2019.03.045. Impact factor (2017): 1.666. Quartile: Q3.
4. J. Diaz-Fortuny, M. Maestro, J. Martin-Martinez, A. Crespo-Yepes, R. Rodriguez, M. Nafria and X. Aymerich, "Current-limiting and ultrafast system for the characterization of Resistive Random Access Memories", AIP Review of Scientific Instruments, vol. 87, no. 6. DOI: 10.1063/1.4954973. Impact factor (2016): 1.515. Quartile: Q3.

Contributions to conferences as first author:

1. J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez, M. Nafria, R. Castro-Lopez, E. Roca, F. V. Fernandez, "CMOS Characterization and

Compact Modelling for Circuit Reliability Simulation” in Proc. 24th IEEE International Symposium on On-Line Testing and Robust System Design (IOLTS), July 2018. **(Invited paper)**.

2. J. Diaz-Fortuny, P. Saraza-Canflanca, A. Toro-Frias, J. Martin-Martinez, E. Roca, R. Rodriguez, F. V. Fernandez and M. Nafria ”A Model Parameter Extraction Methodology Including Time-Dependent Variability for Circuit Reliaility Simulation”, in Proc. 15th International Conference on Synthesis, Modeling, Analysis and simulation Methods and Applications to Circuit Design (SMACD), July 2018. **(Runner-up Best Paper Award)**.
3. J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez, M. Nafria, R. Castro-Lopez, E. Roca, F. V. Fernandez., ”A noise and RTN-removal smart method for parameters extraction of CMOS aging compact models” in Proc. Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS), May 2018.
4. J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez, M. Nafria, R. Castro-Lopez, E. Roca, F.V. Fernandez, E. Barajas, X. Aragones, D. Mateo ”Statistical chatacterization of reliability effects in nanometer CMOS using a versatile transistor array IC”, abstract in the 32nd Conference on Design of Circuits and Integrated Systems (DCIS), November 2017.
5. J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez, M. Nafria, R. Castro-Lopez, E. Roca, F.V. Fernandez, ”Statistical characterization of unreliability effects in a 65-nm CMOS transistor array”, abstract in the 22nd Internatioanl Mixed-Signal Testing Workshop (IMSTW), July 2017.
6. J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez, M. Nafria, R. Castro-Lopez, E. Roca and F. V. Fernandez, ”TARS: A toolbox for statistical reliability modeling of CMOS devices”, in Proc. 14th International Conference on Synthesis, Modeling, Analysis and simulation Methods and Applications to Circuit Design (SMACD), June 2017.
7. J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez, M. Nafria, R. Castro-Lopez, E. Roca, F. V. Fernandez, D. Mateo, E. Barajas, ”A transistor array chip for the statistical characterization of process variability, RTN and BTI/CHC aging”, in Proc. 14th International Conference on Synthesis, Modeling, Analysis and simulation Methods and Applications to Circuit Design (SMACD), June 2017.
8. J. Diaz-Fortuny, M. Maestro, J. Martin-Martinez, A. Crespo-Yepes, R. Rodriguez, M. Nafria and X. Aymerich, ”A new experimental setup for

the ultrafast characterization of the resistive switching phenomenon", in Proc. 19th Workshop on Dielectric in Microelectronics, July 2016.

Co-authored international journal papers:

1. M. Maestro, J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez, M. Nafria, F. Campabadal and X. Aymerich, "New high resolution Random Telegraph Noise (RTN) characterization method for resistive RAM" in Solid-State Electronics (Elsevier), January 2016. DOI: 10.1016/j.sse.2015.08.010, impact factor (JCR 2016): 1.58, quartile: Q3.
2. J. Martin-Martinez, J. Diaz-Fortuny, R. Rodriguez, M. Nafria, and X. Aymerich, "New Weighted Time Lag Method for the Analysis of Random Telegraph Signals" in IEEE Electron Device Letters, April 2014. DOI: 10.1109/LED.2014.2304673, impact factor (JCR 2016): 2.754, quartile: Q1.

Co-authored Conference contributions:

1. 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 Proc. IEEE International Reliability Physics Symposium (IRPS), April 2019.
2. P. Saraza-Canflanca, J. Diaz-Fortuny, R. Castro-Lopez, E. Roca, J. Martin-Martinez, R. Rodriguez, M. Nafria and F. V. Fernandez, "New method for the automated massive characterization of Bias Temperature Instability in CMOS transistors", in Proc. Design, Automation and test in Europe (DATE), March 2019.
3. 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 Proc. 15th International Conference on Synthesis, Modeling, Analysis and simulation Methods and Applications to Circuit Design (SMACD), July 2018.
4. V. M. van Santen, J. Diaz-Fortuny, H. Amrouch, J. Martin Martinez, R. Rodriguez, R. Castro-Lopez, E. Roca, F. V. Fernandez, Jörg Henkel and M. Nafria. "Weighted time lag plot defect parameter extraction and GPU-based BTI modelling for BTI variability" in Proc. IEEE International Reliability Physics Symposium (IRPS), March 2018.

5. C. Couso, J. Diaz-Fortuny, J. Martin-Martinez, M. Porti, R. Rodriguez, M. Nafria, F. V. Fernandez, E. Roca, R. Castro-Lopez, E. Barajas, D. Mateo and X. Aragones, ”Dependence of MOSFETs threshold voltage variability on channel dimensions”, in Proc. Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS), April 2017.

6. M. Maestro, J. Diaz-fortuny, A. Crespo-Yepes, J. Martin-Martinez, R. Rodriguez, M. B. Gonzalez, F. Campabadal, M. Nafria and X. Aymerich, ”Random Telegraph Noise (RTN) analyzed by using Weighted Time Lag Method in Resistive Switching devices” in Proc. 7th International Conference on Unsolved Problems on Noise (UPoN), July 2015.

7. M. Maestro, J. Martin-Martinez, J. Diaz-Fortuny, A. Crespo-Yepes, M. B. Gonzalez, R. Rodriguez, F. Campabadal, M. Nafria, X. Aymerich, Poster: ”Analysis of Set and Reset mechanisms in Ni/HfO_2-based RRAM with fast ramped voltages”, in Proc. 19th Conference on ”Insulating Films on Semiconductors (INFOS), July 2015.

8. M. Maestro, J. Diaz-Fortuny, A. Crespo-Yepes, M. B. Gonzalez, J. Martin-Martinez, R. Rodriguez, M. Nafria, F. Campabadal, X. Aymerich, ”A new high resolution Random Telegraph Noise (RTN) characterization method for resistive RAM”, in Proc. Joint International EUROSOI Workshop and International Conference on Conference on Ultimate Integration on Silicon (EUROSOI-ULIS), January 2015.

9. J. Martin-Martinez, J. Diaz-Fortuny, R. Rodriguez, M. Nafria, X. Ayme rich, E. Roca, F. V. Fernandez and A. Rubio, ”Characterization of Random Telegraph Noise and its impact on reliability of SRAM sense amplifiers”, in Proc. 5th European Workshop on CMOS Variability (VARI), October 2014.

Contents

Agradecimientos i

Acknowledgments iii

Preface v

Publications related to this thesis viii

1 Variability and reliability phenomena in nanometer MOSFETs 1

1.1 Time-zero variability 2

1.1.1 Random Dopant Fluctuations 2

1.1.2 Line-Edge and Line-Width Roughness 3

1.1.3 Oxide Thickness and Poly-Si-Gate granularity 4

1.2 Time-dependent variability: the role of defects in CMOS technologies 5

1.2.1 Transient effects: The Random Telegraph Noise 6

1.2.2 Voltage and temperature dependence of τ_c and τ_e . . . 8

1.3 Time-dependent variability: Aging 9

1.3.1 The Bias Temperature Instability phenomenon 10

1.3.2 The Hot Carrier Injection phenomenon 14
1.3.3 BTI and HCI effects in circuit operation 16
1.3.4 Aging phenomena modelling 17
1.3.5 The Probabilistic Defect Occupancy model 18
1.4 BTI/HCI aging advanced characterization 20
1.4.1 The Enhanced Stress-Measurement aging characterization methodology 21
1.4.2 Defects characterization: The Weighted Time Lag Method 24
1.5 From wafer-level to array-based IC MOSFET characterization 27
1.5.1 Hardware and software design requirements for array-based IC variability characterization 28
1.5.2 Parallel-Stress Pipeline-Measurement method 29
1.5.3 Force-&-Sense voltage biasing 32
1.5.4 Previous works in array-based characterization 33
1.6 Contributions of this thesis 37
2 The ENDURANCE array-based Integrated Circuit chip 39
2.1 ENDURANCE IC design considerations 40
2.1.1 The IC architecture 41
2.1.2 The core: unit cell circuit and layout design 45
2.1.3 Unit cell transmission gates block 49
2.1.4 Unit cell level shifter block 51
2.2 IC digital control: operation modes 53

2.2.1 HCI parallel test example with 3 devices 57

2.3 Conclusions 59

3 A flexible framework for the characterization of TZV and TDV phenomena in array-based ICs 61

3.1 The All DUT Parallel Stress Pipeline Measurements (ADP-SPM) methodology 62

3.1.1 ADPSPM: a time complexity study 66

3.2 TARS software architecture for MOSFET characterization . . 68

3.2.1 Test plan generation module 69

3.2.2 Elementary test instructions generation module 73

3.2.3 Test execution protocol and output test files 75

3.3 Hardware characterization framework 79

3.3.1 Printed circuit board for IC characterization 81

3.3.2 IC leakage current measurement 84

3.4 TARS software: graphical user interfaces 86

3.5 Conclusions 91

4 Expected MOSFET lifetime prediction and parameters extraction methodologies 93

4.1 Lifetime prediction based on time-zero and time-dependent statistical variability measurements 94

4.1.1 Time-zero variability measurements 94

4.1.2 Random Telegraph Noise characterization 100

4.1.3 Bias Temperature Instability characterization 103

4.1.4 Hot Carrier Injection characterization 113

4.1.5 TDV lifetime comparison 117

4.2 Extraction of MOSFET parameters for circuit reliability simulation 119

4.2.1 A model parameter extraction methodology including time-dependent variability 120

4.2.2 A smart noise- and RTN-removal methodology for transistor parameter extraction 126

4.2.3 BTI/HCI-aware circuit simulation using a GPU-based approach 134

4.3 TZV and TDV tests using a 28-nm IMEC IC design 139

4.4 B1530A Ultrafast device characterization 142

4.5 Conclusions 146

Summary and conclusions **151**

Bibliography **151**

CHAPTER 1

Variability and reliability phenomena in nanometer MOSFETs

Imperfections during transistor manufacturing process, causing variations of the fabricated devices, worsen with technology downscaling. Shifts of intrinsic transistor parameters, like threshold voltage (V_{th}) or mobility (μ) at the fabrication process, as well as their progressive variation during circuit operation over time, are major culprits behind fatal performance deviations that can negatively affect fabricated integrated circuits (IC) design and must be statistically characterized and modelled [1,2]. The combination of time-zero variability (TZV) and time-dependent variability (TDV) phenomena affects and compromises the reliability of the fabricated devices and circuits reducing, for instance, their intended lifetime.

In this scenario, the present chapter will briefly describe the most relevant sources of TZV and TDV phenomena affecting nowadays nanometer MOSFETs. Moreover, the methodologies utilized to experimentally characterize devices under accelerated aging tests will be described. In the last sections of this chapter, a comparative analysis between conventional wafer-level and array-based IC transistor characterization is presented. Moreover, the hardware and software design requirements for a trustworthy array-based IC device characterization under aging phenomena will be described. In this sense, an analysis of the fulfilment of these demands in the most relevant array-based IC works presented in the recent literature is included. Finally, the objectives and contributions of this thesis are presented.

1.1 Time-zero variability

Time-zero variability (TZV), also known as process variability, consists in permanent deviations of the transistor parameters from their intended nominal values that occurs during the fabrication process. Typically, TZV is divided into stochastic and systematic variability. The stochastic group is defined as the local or intra-die process variability and the systematic component is defined as the global or inter-die process variability. On the one hand, local process variability causes parametric variations or mismatch between identically designed devices within the same die. On the other hand, global process variability causes die-to-die, wafer-to-wafer, or lot-to-lot variations between identical designed devices [3–5].

The main sources of TZV affecting CMOS device related to CMOS transistor manufacturing are: Random Dopant Fluctuations (RDF), Line-Edge Roughness (LER) and Line-Width Roughness (LWR) variations, Oxide Thickness Variations (OTV) effects and Poly-Si-Gate granularity. The randomness of the mentioned atomistic effects is uncorrelated from device-to-device converting each fabricated transistor in a unique device in terms of electrical parameters [6, 7].

1.1.1 Random Dopant Fluctuations

Random dopant fluctuations (RDF) are caused by stochastic variations in the number and location of the dopant atoms in the channel region of a MOSFET transistor. These dopant atoms are in charge of controlling, for example, the V_{th} of the fabricated devices. Thus, each transistor will have a unique doping profile that causes random variations in the transistor parameters like the V_{th} or the on-state current I_{ON}. Figure 1.1 shows a representative example where the major variability sources in a MOSFET device can be located. The RDF profile can be easily located inside the substrate region of the MOSFET (see Figure 1.1) where a random number of blue dots, featuring the Random Discrete Dopants, produces a unique atom doping profile [7, 8].

The effect of RDF increases with technology scaling. This is because the average number of dopant atoms in the transistor channel decreases exponentially with technology scaling. For instance, assuming a doping density of $10^{18}/cm^3$, the average number of dopant atoms in a hypothetic

channel with Width (W) of 70nm and Length (L) of 40nm is approximately 100. Thus, the variation of the electrical parameters between identically fabricated MOSFET devices depends on the low and discrete number of dopant atoms in devices [9].

In this scenario, it is necessary to statistically characterize and model the effects of RDF at the device and circuit levels to enhance reliability of digital and analog circuits in nanometer scale regimes. For instance, the mismatch produced by the RDF in the two cross-coupled inverters of a basic SRAM cell can result in significant variation in the delay of a the circuit, data loss or bit flip [8, 9].

1.1.2 Line-Edge and Line-Width Roughness

Other important source of transistor variability related to nanometric channel dimensions are the Line-Edge Roughness (LER) and Line-Width Roughness (LWR) effects. Both phenomena consist in imperfections or variations in the gate edges of devices that result in non-rectilinear shapes with respect to a perfect linear edge. Figure 1.1, shows an illustrative example of both LER and LWR effects in the gate region of a MOSFET device. The source of these imperfections is related to the lithography and etching processes during transistor fabrication [10, 11].

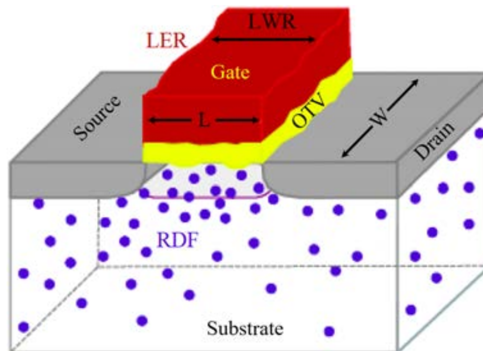


Figure 1.1: Example of LER/LWR, RDF and OTV in a MOSFET device [3].

LER and LWR effects do not scale according to the CMOS technology scale trend. In large transistors (channel length above 100nm) LER and LWR

were not a relevant source of variability since the edge roughness caused at the fabrication stage remains typically in the order of ≈5-nm independently of the fabrication process. However, in current planar MOSFET, FinFET and multi-gate fabrication processes, at the sub-100-nm technology nodes, the contribution of LER/LWR is becoming a comparable fraction of the transistor gate L and W thus, the resulting variability is significantly increased [12, 13]. Moreover, LER/LWR variability effects on critical device parameters have revealed that the inclusion of this phenomena in variability simulations have resulted in V_{th} and I_{ON}/I_{OFF} variations close to the nominal distributions of the transistor parameters [14, 15].

Studies where RDF and LER/LWR effects have been simulated together, have concluded that the combined phenomena effects are statistically independent. Moreover, the effects of LER/LWR in the device parameters variations have a much more stronger effect for channel length below 20-nm than for devices with higher channel lengths where, the RDF become the dominant TZV source [12, 14, 15].

1.1.3 Oxide Thickness and Poly-Si-Gate granularity

In advanced technology nodes, device insulator thickness below 10nm are responsible for an important source of intrinsic parameter fluctuations such as the oxide thickness variations (OTV) and Poly-Si-Gate granularity.

OTV are caused by imperfections or roughness in the silicon/insulator and the gate/insulator interfaces across the gate stack of transistors as shown in Figure 1.2 (a). The OTV cause variations in the V_{th}, μ [16] or gate current tunnelling [17] between identical fabricated devices.

The use of highly doped polycrystalline granular structures in the fabrication of Poly-Si gates of MOSFETs to obtain electrical properties close to metals, in combination with thin gate oxides, introduce another source of transistor parameter variability known as Poly-Si-Gate granularity. The Poly-Si-Gate granularity variability occurs when gate dopants do not diffuse optimally down to the gate/insulator interface. Then preferential dopant diffusion along grain boundaries results in non-uniform isolated doped regions in the transistor gate interface generating local gate depletion zones as shown in Figure 1.2 (b). This phenomenon results in an effective oxide thickness increase and introduces additional random dopant charge on the side of the

gate oxide opposite to the channel enhancing, for instance, V_{th} fluctuations from device to device [18]. Moreover, the introduction of High-k materials in the transistor insulator with the aim of reducing leakage currents due to technology scaling, also leads to localised variations in gate-to-channel interface potential that results in V_{th} variations between devices [19–21].

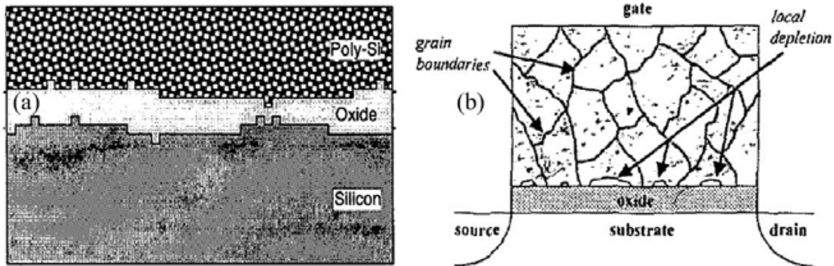


Figure 1.2: Oxide Thickness Variations in a MOSFET device in (a) and Poly-Si-Gate insulator granularity variability in (b) [16] showing Poly-Si gate boundaries and local gate depletion zones [18].

1.2 Time-dependent variability: the role of defects in CMOS technologies

One of the reasons for the increase of transistor TZV and TDV is the reduction of the transistor's gate area, that must be linked to gate insulator thickness (T_{OX}) scaling, to mitigate the short channel effects of nanometric devices. It is well known that conventional thickness scaling of traditional SiO_2 insulators became unfeasible at the 90nm technology node and below because of the intolerable high gate leakage current mainly produced by direct tunnelling through the gate insulator [22, 23].

In this scenario, the semiconductor industry has moved forward and started to manufacture transistors first utilizing Silicon Oxynitride (SiON) gate insulators. Later on, the introduction of High-K Metal gate insulators (HKMG) at the 45-nm and 28-nm node allowed the continued scaling of the device geometry without a significant increase of the gate leakage currents. Nevertheless, to control the increasing short channel effects at the 20-nm, 16-nm and 14-nm technology nodes, the industry has migrated from planar MOSFET to FinFET, FD-SOI FETs or Multi-Gate FETs [24–26].

With the rapidly CMOS scaling trend imposed to newer CMOS technology nodes, the influence of the previously described sources of device TZV in the shift of the intrinsic transistor parameters, must be completed with device transients and degradation phenomena during its operation over time. In this sense, the Random Telegraph Noise (RTN) transient phenomenon and the Bias Temperature Instability (BTI) and the Hot-Carrier Injection (HCI) aging phenomena [27, 28] have become critical issues in current ultrascaled technologies. These mechanisms consist in the stochastic charge trapping/detrapping in/from device defects located at the device insulator, bulk or at the interface. Moreover, the defects present in a CMOS device can be inherent to the manufacturing process or also generated during transistor operation as a consequence of BTI/HCI aging [29].

Defect charge trapping/detrapping may cause significant transistor parameters variations, for instance V_{th} or μ, to shift from their as-fabricated values that can cause device mal-function or a permanent failure [30]. To account for the variability produced by the charge trapping/detrapping, three main parameters of individual defects should be statistically characterized: the capture time (τ_c) (the mean time needed to capture a charge by a defect), the emission time (τ_e) (the mean time needed to emit a charge by a defect) and the drain current shift (ΔI_{DS}) or, equivalently, the threshold voltage shift (ΔV_{th}) associated to each defect, typically known as eta (η). The analysis of the three defect parameters revealed the following characteristics:

1. Both capture and emission time constants (τ_c, τ_e) are uncorrelated and bias and temperature dependent [31, 32].
2. Capture and emission defect time constants, can expand several decades in time thus, the selection of an appropriate measurement time window to characterize them is of high importance [33].

1.2.1 Transient effects: The Random Telegraph Noise

The Random Telegraph Noise (RTN) is characterized by a random switching of the I_{DS} between two or more levels along time and has been recognized as a significant variability source since it is responsible for device parameter variability, such as the ΔV_{th} [33] at nominal operation conditions. These current fluctuations are related to the capture and emission of charge carriers by oxide and interface defects, showing a large dependence on device biasing and temperature conditions. Moreover, RTN variability phenomenon increases

inversely with area scaling negatively affecting, for instance, analog devices and digital logic circuits like flash memories due to the V_{th} variability [34, 35]. Moreover, RTN is not only present in conventional MOSFET transistors, it can be detected in Resistive Random Access Memories (RRAM) utilized in the design of non-volatile memories [36], Fully Depleted FET (FD-SOI) [37], FinFET [38], Multigate FET [39, 40], and nanowire FET [41] devices and digital circuits [42].

Standard RTN characterization, that consists of executing a continuous measurement of the I_{DS} while maintaining a constant V_{GS} and V_{DS} biases, permits the studying of the RTN active defects parameters, determine its location in the device and the defect energy depth below the conduction band of the dielectric [33, 43]. An accurate measurement is required since the current fluctuations can vary just a few nanoamps while the time required for the RTN characterization depends on the defect intrinsic capture/emission time constants, which may span several orders of magnitude in measurement time. For instance, Figure 1.3 shows two examples of typical RTN signals where the I_{DS} fluctuate between two different levels that correspond to the charge/discharge of a single active defect in a measurement time window of 2s each.

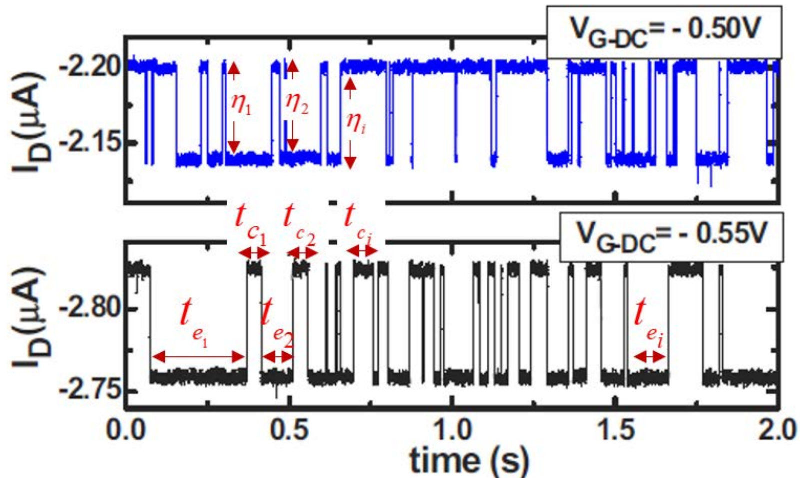


Figure 1.3: Example of two RTN signals showing the location of individual emission times t_{e_i}, capture times t_{c_i} and the current fluctuations amplitude (η_i) [44].

In this scenario, the RTN fluctuations originated from a single defect can be fully characterized by three parameters: η, that describes the average impact on the V_{th} or I_{DS} amplitude fluctuation, the τ_c as the mean elapsed time before the system captures a charge carrier ($< t_{c_i} >$) and τ_e as the mean elapsed time before the system emits a charge carrier ($< t_{e_i} >$) (see Figure 1.3 for individual t_{c_i} and t_{e_i} occurrence) [45].

1.2.2 Voltage and temperature dependence of τ_c and τ_e

As already introduced in section 1.2, charging and discharging of device defects may produce I_{DS} device variations that are directly related to V_{th} fluctuations. This behaviour that compromises device reliability become the source of transient effects like RTN as well as TDV aging phenomena such as BTI and HCI [32]. In this sense, dependence with voltage and temperature conditions of the t_e, t_c times associated to those variations [32, 46, 47] will be shown. For instance, Figure 1.4 (a) and (b) shows the statistical distribution of t_e and t_c respectively, of an individual device defect from an RTN measurement with fixed V_{DS} and increasing V_{GS} conditions. In both Figures 1.4 (a) and (b), t_{c_i} and t_{e_i} parameters have been plotted assuming its exponential distribution where F, the accumulated distribution function, can be expressed by equation 1.1.

$$ln\left(1 - F\left(V, T\right)\right) = \frac{-t_{c,e}}{< t_{c,e}\left(V, T\right) >} \tag{1.1}$$

As can be extracted from Figure 1.4 (a), for increasing V_{GS} voltage, the slope of the t_e distribution decreases, which means that the emission times t_e increase. On the contrary, the slope of the t_c in Figure 1.4 (b) increases with increasing V_{GS} meaning that the capture time t_c decreases [48].

Recent works have demonstrated that defect t_c and t_e times are reduced with high temperature revealing that the probability to charge or discharge a defect increases with temperature. For instance, Figure 1.5 (a) and (b) shows the statistical distributions of the t_e and t_c, respectively, of a single defect showing that a temperature increase produces a defect time constant reduction in both cases.

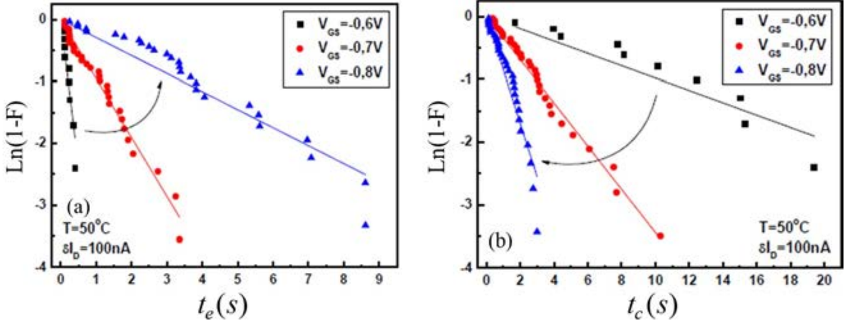


Figure 1.4: Statistical distributions of (a) t_e and (b) t_c for a $\Delta I_{DS} = 50nA$ criterion and a fixed temperature of $T = 50°C$ at different V_{GS} device biasing. Lines correspond to the experimental fitting using equation 1.1 where the mean values of $< t_{e_i} >$ and $< t_{c_i} >$ can be obtained as the inverse of the fitting slopes respectively [49].

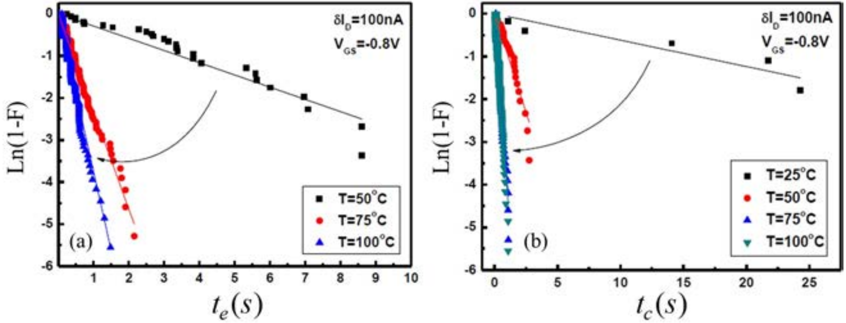


Figure 1.5: Statistical distributions of the (a) t_e and (b) t_c of a single device defect that produces a $\Delta I_{DS} \approx 100nA$ for different temperatures [49]

1.3 Time-dependent variability: Aging

A progressive degradation of the intrinsic device parameters takes place after long periods of time (months or years) working at nominal operation conditions due to aging mechanisms like the Bias Temperature Instabilities and the Hot Carrier Injection phenomena. These are sources of TDV affecting the transistor parameters over time. In this scenario, it becomes practically unfeasible to characterize BTI/HCI TDV aging at the nominal operation conditions such long times.

Typical aging characterization methods are based in the execution of accelerated aging tests, in which temperature and/or the drain and/or the gate voltages are raised above their nominal operation values over a much shorter period of time. Thus, transistor parameters degradation is produced in significantly reduced time. High voltages and/or temperatures are referred to as the stress conditions while, after the stress, devices are measured close to the nominal operation conditions to compute transistor parameter shifts, if present, without introducing significant damage to the device. For instance, degradation of V_{th} is one of the essential parameters to determine MOSFET degradation metrics and serves as an effective indicator when evaluating device variability and reliability [23].

1.3.1 The Bias Temperature Instability phenomenon

The Bias Temperature Instability is one of the most relevant reliability concerns for the nanometric semiconductor industry known by academia and industry for more than 50 years. BTI degradation is known for pMOS devices as Negative BTI (NBTI) and occurs when the gate of a transistor is negatively biased resulting in a build-up of positive charge in the gate insulator. NBTI phenomena progressive degrades intrinsic device parameters such as the linear (ΔI_{DLIN}) and saturation (ΔI_{DSAT}) drain current, the threshold voltage (ΔV_{th}), transconductance (Δgm), among others. For the nMOS devices, the BTI is known as Positive BTI (PBTI) and appears when a positive bias is applied to the gate of the device while all the other terminals are grounded. PBTI causes an accumulation of negative charge in the gate insulator that results in a positive shift of, for instance, the device V_{th} [50, 51].

NBTI was first observed in devices based on Silicon Dioxide (SiO_2) insulators. Nevertheless, it also continues as an important concern with the introduction of new insulators like Silicon Oxynitride (SiON) [52] around the year 2000 [53] where NBTI was the predominant source of transistor degradation. PBTI has remained less predominant for nMOSFETs with SiO_2 and SiON gate stacks until the introduction of newer High-K Metal Gate dielectrics (HKMG) [54] and advance devices such as FinFETs [55, 56].

An intrinsic feature of the BTI phenomena, caused by the charge and discharge of device defects, is that after the application of stress, where device defects can be charged, the shift of the absolute positive parameters, like ΔV_{th} or $\Delta\mu$, exhibit a recovery behaviour that tends to reach its non-degraded

values immediately after the stress removal. This process is also known as BTI recovery or relaxation and can be prolonged for several decades in terms of measurement time. For instance, Figure 1.6 shows an example of the described stress-recovery BTI aging behaviour where the ΔV_{th} parameter is shown for pMOS and nMOS HKMG devices in Figure 1.6 (a) and (b) respectively [55]. A $|\Delta V_{th}|$ occurs for both pMOS and nMOS devices during the stress phase under BTI aging.

After the removal of the stress conditions, $|\Delta V_{th}|$ starts recovering its original value due to the discharge of previously charged defects [57]. It is important to remark that for the relaxation window of 1000s shown in Figure 1.6 (a) and (b), the total ΔV_{th} parameter does not reach a total recovery, i.e., $\Delta V_{th} = 0V$ and a plateau appears as the BTI permanent component (P) [58, 59].

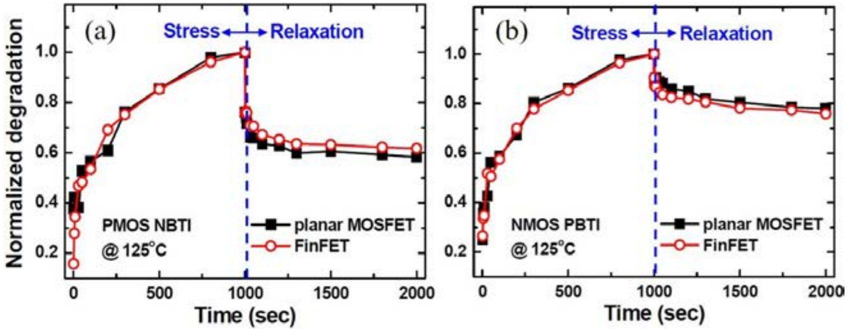


Figure 1.6: Example of the ΔV_{th} shift behaviour during BTI stress and relaxation processes for 20-nm planar and 14-nm FinFET devices for pMOS devices in (a) and for nMOS devices in (b) [55].

For long channel devices, i.e., channel lengths in the micrometer scale, BTI behaviour can be detected with a smooth recovery trend due the contribution of many device defect discharges as shown in Figure 1.7 (a). On the contrary, in modern deep scaled devices, i.e., channel lengths in the nanometer scale, BTI recovery behaves in discrete steps, as shown in Figure 1.7 (b), where abrupt $-\Delta V_{th}$ defect discharges can be easily detected because only a few defects are present in the device and their contribution is much more severe. Moreover, BTI recovery starts immediately just after the stress biasing conditions are removed. Thus, ΔV_{th} degradation is very sensitive to any measurement delay where, the higher the measurement delay after stress removal the lower the ΔV_{th} can be captured.

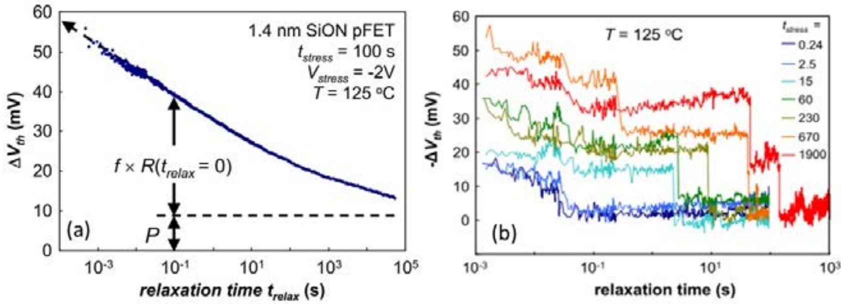


Figure 1.7: (a) smooth ΔV_{th} BTI recovery trace obtained from a pFET device with SiON dielectric and L between $0.13\mu m$ and 0.5 μm and W = $10\mu m$ dimensions, after BTI stress of V_{GS} =-2V [60]. (b) 7 BTI recovery traces showing discrete $-\Delta V_{th}$ variations due to defect discharges for a pFET device (L = 70nm, W = 90nm) and HfO_2 based dielectric [50]

.

Defects responsible for BTI degradation are voltage and temperature dependent as well as frequency dependent. From a temperature point of view, BTI degradation increases for elevated temperatures. Moreover, defects that contribute to BTI degradation, show individual activation energy E_A for capture and emission time constants following the Arrhenius law [61]. In this regard, Figure 1.8 (a) shows an example of the temperature dependence of both the recoverable and the permanent BTI components when increasing the temperature from 25°C to 200°C revealing the temperature dependence of the $< \Delta V_{th} >$ parameter. Newer NBTI characterization techniques have emerged in recent literature maximizing the degradation/relaxation utilizing very high stress temperatures by means of controllable poly-heaters [62] to assess the long-term BTI phenomena in the shortest time possible [63, 64].

In terms of BTI frequency dependence, several works have been published in the literature reporting AC-BTI weak frequency dependence [65, 66] while other studies have demonstrated a frequency-dependent contribution of BTI [67,68] which exemplifies that the AC-BTI dependence is subject of controversy. For instance, Figure 1.8 (b) shows a comparison of the recovery behaviour between a previous DC stress of 1ks with $V_{G-stress}$ = -2V and $V_{G-relax}$ = -0.3V (open circles) and two AC stress waveforms during 2ks with f_{AC1} = 1Hz (blue squares) and f_{AC2} = 100kHz (red squares). If the AC on/off waveforms oscillate between $V_{G-stress}/V_{G-relax}$ (-2V/-0.3V), the ΔV_{th} for DC and AC stress merges only at the end of the recovery window while at the beginning circles and squares are clearly separated. This behaviour demonstrates a

frequency dependence of the BTI recoverable component while the permanent component is not. Nevertheless, if a more positive AC-off value is used, (-2V/+0.5 and -2V/+1V), the frequency dependence is mitigated (see diamonds and triangles in Figure 1.8 (b)) [67].

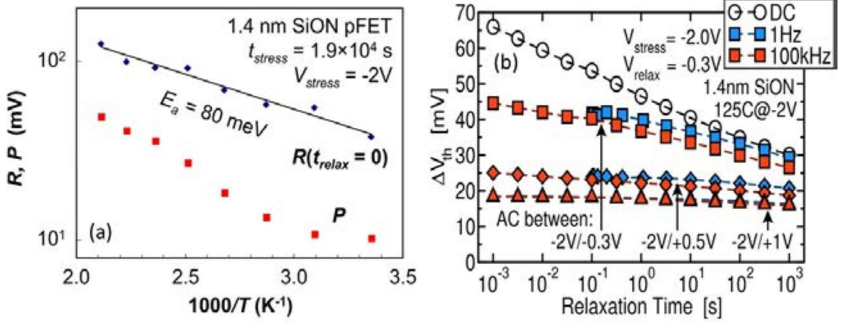


Figure 1.8: Example of the temperature dependence of BTI phenomena of both the recoverable (R) and the permanent (P) component under DC stress in (a) [60] while (b) shows the BTI aging frequency dependence due to AC stress [67]

The following considerations about BTI phenomena behaviour are well accepted [32, 69–71]:

1. Even for short stress times ($< 1s$), emission and capture time events show a wide range of characteristic time scales, ranging from μs time scale to several decades in terms of logarithmic experimental window.
2. Both τ_c and τ_e times of each defect are determined by its energy level (E_T), the defect depth into the gate stack and gate bias (V_{GS}). Moreover, the capture and emission events of each defect are stochastic and exponentially distributed.
3. BTI defects are considered the same defects that originates the RTN switching phenomenon. Thus, the difference between BTI and RTN phenomena defect behaviour is that in an RTN experiment, a limited number of defects with capture and emission time constants within the experimental window are visible. On the contrary, due to the bias dependence of the capture time constant, more defects contribute to BTI because of a previous device stress that allows charging more defects than in an RTN experiment.

1.3.2 The Hot Carrier Injection phenomenon

The HCI aging phenomenon has been known for nearly five decades and consists of a degradation mechanism where carriers along the transistor channel are accelerated because of the imposed lateral electric field, $(V_{GS}, V_{DS}) > 0V$. This makes the carriers to gain sufficient energy to create insulator defects altering permanently the electrical properties of the device [72, 73].

Early in the 1970s-1980s, the positive supply voltage (VDD) in the CMOS industry was fixed while transistor area began to scale, resulting in a strong increase of the device vertical/lateral electric fields and thus, turning HCI aging the most relevant transistor reliability issue. In the 1990s, VDD was reduced with each technology node as well as the electric fields, attenuating the HCI effects in the device/circuit reliability. Nevertheless, nowadays below the 100-nm node, the VDD biasing values are saturating around $\approx$ 1V, increasing the vertical and lateral transistor electric fields and, therefore, HCI aging phenomena is becoming a more severe aging mechanism again [74–76].

The HCI degradation can be observed as a quasi-permanent shift in the device intrinsic parameter such as V_{th}, a reduction of the saturation current I_{DSAT}, between others. In contrast to nanometer-scale devices where the worse bias combination for HCI damage is $V_{GS} = V_{DS}$, early in the 1980s the critical biasing condition used to occur at $V_{GS} = V_{DS}/2$ [77–79]. In this sense, Figure 1.9 shows an example of the ΔV_{th} degradation suffered by planar nMOS devices with L = 27nm and W = 90nm. As shown in Figure 1.9, ΔV_{th} degradation under stress conditions reveal high degradation for increasing stress voltage.

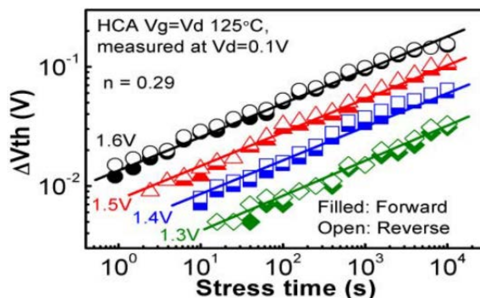


Figure 1.9: HCI degradation showing ΔV_{th} shift after 4 $V_{GS} = V_{DS}$ stress voltages and measurements carried out at $V_{DS} = 0.1V$, $V_{GS} \approx V_{th}$ [78].

HCI phenomenon is also frequency and temperature dependent. In this sense, Figure 1.10 (a) and (b) shows two examples of device parameter degradation due to HCI accelerated degradation under different temperature conditions and AC/DC stress respectively. Figure 1.10 (a) depicts the ΔV_{th} degradation for various V_{DS}/T stress conditions showing that under high temperature (T=398K), HCI degradation is larger than at low temperatures (T=298K). On the other hand, Figure 1.10(b) shows the ΔI_{DSAT} at two different temperatures under AC and DC stress conditions. The results reveal a lower degradation when stressing with AC waveforms for the two temperatures than for DC stress [80].

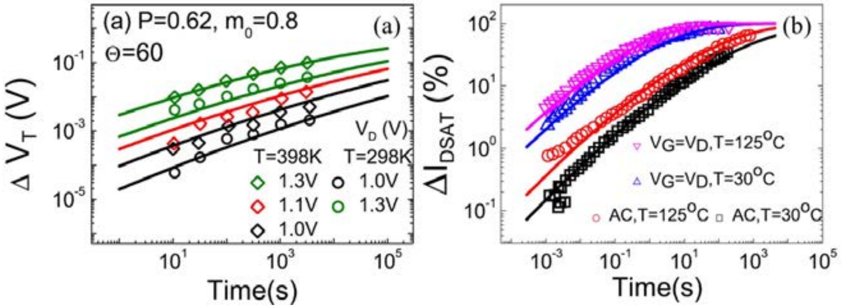


Figure 1.10: Intrinsic device parameters degradation under HCI aging phenomenon showing in (a) the ΔV_{th} for different V_{DS} and temperature conditions and in (b) the ΔI_{DS} for AC and DC stress at different temperatures [80].

To describe HCI degradation effects, the traditional Lucky Electron model (LEM) [81] based on direct electron excitation was widely accepted for long channel devices. As LEM postulates, the cause for HCI degradation involves defects generated by electrons (e^-) or holes (h^+) that impact into the device interface thanks to the high electric lateral fields in the device.

For nanometric devices, the LEM model predicts a very low effect of the HCI phenomena because electric fields are reduced due to low supply voltages. Nevertheless, recent literature shows that HCI degradation remains significant for ultra-scaled MOSFETs and therefore new models are needed. In contrast to the conventional field-driven LEM, recent literature proposed novel energy-driven theories [82, 83] to describe HCI phenomenon and accurately explain and model the mechanism of carrier induced degradation and to overcome the limitation imposed by the traditional LEM [79, 82, 84].

1.3.3 BTI and HCI effects in circuit operation

In real circuit operation, BTI and HCI aging degradation can coexist degrading devices and circuits while reducing the intended reliability of the overall circuit. In this sense, Figure 1.11 shows an explanatory example where a simple CMOS inverter designed with a pMOS and a nMOS transistor, connected in series between VDD and VSS biased using a rail-to-rail, i.e., from VDD to VSS, square signal. The following situations can occur during circuit functionality:

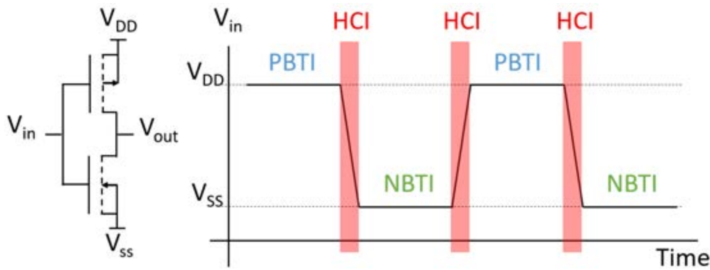


Figure 1.11: Schematic representation of a CMOS inverter under a typical square input signal showing BTI and HCI reliability phenomena.

1. Logical '1' input (V_{in}= VDD): the nMOS transistor is in on-state because $V_{GS} = V_{GB}$ =VDD, while the pMOS transistor remains in off-state due to $V_{GS} = V_{GB} = 0V$. In this scenario, the PBTI phenomena appears in the nMOS transistor (PBTI blue label in Figure 1.11) degrading the nMOS device. Moreover, the pMOS transistor experiences the so called Off State Stress (OSS) [85, 86] because of the high $|V_{DS}|$ =VDD while $|V_{GS}| = 0V$.
2. Logical '0' input (V_{in} = VSS): in this case, the reverse situation can be observed in the inverter. NBTI phenomenon appears in the pMOS transistor because of the $|V_{GS}| = |V_{GB}|$ =VDD biasing (NBTI green label in Figure 1.11), and the nMOS transistor suffers from OSS because of the high lateral electric field when V_{DS} =VDD and V_{GS} = 0V.
3. HCI device stress occurs during the dynamic switching periods of the square signal when the V_{GS} and the V_{DS} voltages are high enough and there is current flowing through the devices. In this situation, when the input signal changes from VDD to VSS or vice versa, HCI degradation appears in the nMOS and the pMOS transistors (se red rectangles in Figure 1.11).

The impact of TDV phenomena including transient and aging effects in transistor and circuit reliability, have achieved an important relevance in current advanced technology nodes. TDV aging modelling techniques and advanced characterization methodologies have been proposed in the literature to understand and mitigate degradation effects in the fabricated transistors that will be briefly described in the next sections of this chapter.

1.3.4 Aging phenomena modelling

To accurately model the effects of aging phenomena related to BTI/HCI aging mechanisms, several physical models have been proposed in the literature to extrapolate to the nominal operation voltages, the accelerated degradation suffered by devices under aging tests. Nevertheless, experimental results have demonstrated that initial models, like the original Reaction-Diffusion model [87], that was latter improved by the works in [88–90], are not suitable to completely describe the aging phenomena in nanometer devices [50, 91]. In this sense, newer models can be found in the literature to describe the defect behaviour during BTI/HCI and RTN trapping/detrapping related phenomena: the Two-Stage Model [92, 93], the Two-energy-well model [94], the Non-radiative two-stage Multi-Phonon (NMP) model [95] or the four-state NMP model [96].

Despite the fact that NMP models provide an accurate physical description of aging variability phenomena, they typically require computational intense TCAD simulations with numerous physical model parameters. In this sense, reaction-limited models such as the capture/emission time map (CET) model [97] or the Probabilistic Defect Occupancy (PDO) model [98], which will be described hereinafter, were proposed in recent literature as efficient alternatives because they use a limited number of parameters and provide a physical approximation for the NMP models [95].

In the context of accurate device and circuit modelling, this thesis will provide statistically accurate parameters to feed the PDO model because its simplicity to be used with experimental device measurement without scarfing accuracy when reproducing aging degradation compared with other models [99].

1.3.5 The Probabilistic Defect Occupancy model

The PDO model allows describing the V_{th} variability produced by BTI and HCI aging phenomena. Moreover, the model permits obtaining V_{th} variability under arbitrary V_{GS} biasing conditions and is based on the stochastic properties of the individual defects localized in MOSFET devices. It assumes that each individual device has a finite number of active defects (N) that can be charged or discharged. This behaviour will produce different V_{th} variations in the device due to the contribution of each defect. In this scenario, an active defect can capture a charge producing a decrease of the I_{DS} device current or if a defect releases a charge, an increase of the I_{DS} current is observed [52].

The PDO model permits the evaluation of the ΔV_{th} by analysing the state (charged or discharged) of each defect as defined by equation 1.2. For a given stress time t_s and recovery time t_r considering (i) the defect distribution ($D(\tau_e, \tau_c)$) of the device in the (τ_e, τ_c) space and (ii) the occupancy probability of a defect, P_{occ}, defined by the probability if a defect is charge or discharged. The P_{occ} depends on the MOSFET device operation conditions and the (τ_e, τ_c) values, the number of defects (N) and the mean defect impact in the $< \Delta V_{th} >$ (η)

$$\Delta V_{th}\left(t_s, t_r\right) = N < \eta > \int_0^{\infty} \int_0^{\infty} D\left(\tau_e, \tau_c\right) \cdot P_{occ}\left(\tau_e, \tau_c, t_s, t_r\right) \partial\tau_e \partial\tau_c + P_p \tag{1.2}$$

By using the PDO model a log-normal bivariate distribution is utilized for simplicity, as the device defect distribution, which is defined by the mean values of τ_c and τ_e times, $< \tau_c >$ and $< \tau_e >$, their standard deviations σ_{τ_c} and σ_{τ_e} and a correlation coefficient ρ between σ_{τ_c} and σ_{τ_e}. The $D(\tau_e, \tau_c)$ can be analytically obtained with the equation 1.3.

$$D\left(\tau_e, \tau_c\right) = \frac{1}{2\pi\sigma_{<\tau_c>}\sigma_{<\tau_e>}\sqrt{1-\rho^2}} \cdot exp\left[-\frac{1}{2\left(1-\rho^2\right)} \cdot \right.$$
$$\left.\left[\left(\frac{\tau_c - <\tau_c>}{\sigma_{<\tau_c>}}\right)^2 \left(\frac{\tau_e - <\tau_e>}{\sigma_{<\tau_e>}}\right)^2 - 2\rho\left(\frac{\tau_c - <\tau_c>}{\sigma_{<\tau_c>}}\right)\left(\frac{\tau_e - <\tau_e>}{\sigma_{<\tau_e>}}\right)\right]\right] \tag{1.3}$$

For instance, Figure 1.12 (a) shows the occupancy probability (P_{occ}) together with an example of a 2-dimension representation of a particular defect distribution $D(\tau_e, \tau_c)$ in the (τ_e, τ_c) space analytically calculated with equation 1.3 for an nMOS transistor.

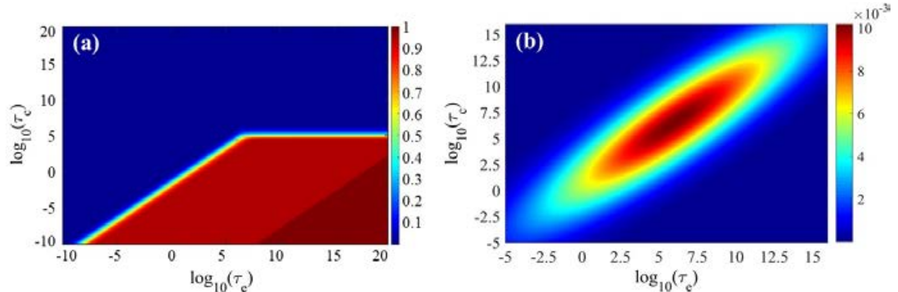


Figure 1.12: (a) shows an example of the occupancy probability as a function of τ_c and τ_e at DC stress [98], and (b) shows an example of the defect distribution $D(\tau_e, \tau_c)$ graph at V_{GS} = 1.2V, V_{DS} = 0V at 25°C.

Using the proposed distribution of defects $D(\tau_e, \tau_c)$ and the occupancy probability P_{occ}, the ΔV_{th} induced variability can be determined for a particular transistor size and a particular stress time (t_s) and recovery time (t_r). All defect transitions captured inside the product of $D(\tau_e, \tau_c)$ and P_{occ} will contribute to the ΔV_{th} for the experimental t_s and t_r. For instance, Figure 1.13 (a) shows an example of BTI measured recovery traces and the resulting PDO simulated traces in ΔV_{th} in Figure 1.13 (b) on a pMOS transistor with a single stress/recovery aging cycle.

An important consideration regarding the $D(\tau_e, \tau_c)$ shown in Figure 1.12 (b) must be extracted. The (τ_c, τ_e) space expands several orders of magnitude in time that cannot be fully covered experimentally. Conventional experimental windows used to characterize aging phenomena cover up to $log_{10}(\tau_e)$ from 0 to 4 and $log_{10}(\tau_c)$ from -3 to 2, thus, the experimental achieved portion of the $D(\tau_e, \tau_c)$ is small. To overcome this limitation, large number of devices should be measured to achieve sufficient statistical data to reproduce the complete defect distribution. In this sense, the next section will describe the typical characterization procedures for aging testing introducing the new Enhanced Stress-Measurement methodology developed in this thesis for accurate BTI/HCI aging characterization.

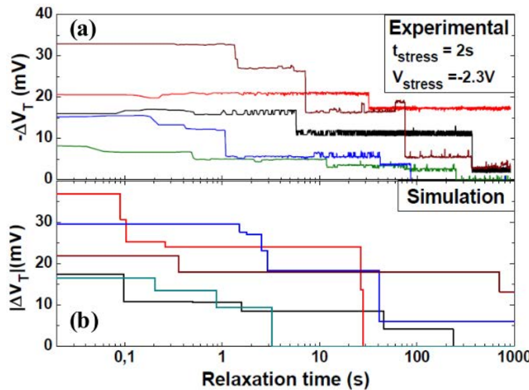


Figure 1.13: (a) Experimental recovery traces in pFETs with abrupt defect discharge events. (b) Simulated recovery traces obtained using the PDO model using the same conditions of stress and recovery than in (a) [98].

1.4 BTI/HCI aging advanced characterization

To obtain the electrical transistor parameter shifts under BTI/HCI aging degradation, the Measurement-Stress-Measurement (MSM) technique has become the most widely used methodology for device characterization. The MSM technique consists in concatenating a device pre-stress measurement period, followed by a stress period and, finally, a device post-stress measurement period. To accelerate aging process, an over-than-nominal V_{GS} and/or V_{DS} voltages should be applied, e.g., 2x or 3x the technology VDD. Moreover, during the pre- and post-stress measurement periods, I_{DS}–V_{GS} curve characteristics are commonly used to account for the device degradation. Finally, the MSM sequence is repeated sequentially, increasing the stress time to further extrapolate the device degradation to the nominal operation conditions. For instance, the stress time increase ratio can be defined exponentially (1s, 10s, 1,000s, 10,000s...) but this increase rate can be of any kind depending on the research needs.

Nevertheless, for BTI tests, the conventional MSM technique suffers from unwanted loss of measurement data, because a big portion of the accelerated degradation during stress starts recovering, without being measured, towards their non-stressed values immediately after the stress removal [100, 101]. To try to overcome this issue, several techniques have been proposed in the literature for aging characterization: the ultra-fast MSM (UF-MSM) [102],

the Ultra-Fast On-the-Fly (UF-OTF) [103, 104], and the One Spot Drop Down (OSDD) methods [23].

The main issues related to the conventional MSM and UF-MSM characterization techniques is that they are prone to suffer from initial recovery data loss because of the time required to compute an I_{DS}–V_{GS} curve after the stress. Moreover, the use of I_{DS}–V_{GS} curves sweeps can mask the step-like behaviour of the recovery trend (see Figure 1.7).

In the UF-OTF case, the I_{DLIN} degradation is measured utilizing a small gate pulse around the $V_{G-STRESS}$ value to account for the transconductance (g_m) change that results in $\Delta V_{th} = \Delta I_{DLIN}/g_m$. The main disadvantage of this technique is that it assumes that the total degradation in the I_{DLIN} is caused only by V_{th} degradation without incorporating the degradation of the μ which, in fact, suffers from degradation, what can impact the accuracy of the method [105]. Finally, the OSDD methodology where the $I_{DS} - t$ is characterized at the nominal biasing conditions after stress has become one of the best candidates to characterize the step-like behaviour of the measurement periods in aging tests. Nevertheless, the OSDD technique is prone to suffer from unwanted recovery data loss because of the time gap required to change from stress to measurement biasing conditions when measuring at wafer level. In this regard, the present section will describe in detail an experimental aging characterization methodology entirely developed in this thesis based on the classical MSM and OSDD methodologies enhanced to minimize recovery data loss and increasing accuracy of the measurement periods after stress.

1.4.1 The Enhanced Stress-Measurement aging characterization methodology

The Enhanced SM method (eSM), based in the MSM and an the OSDD test techniques, utilizes a unique I_{DS}–V_{GS} initial measurement period to account for the non-degraded device parameters followed by multiple stress-measurement (SM) cycles executed sequentially in order to account for the accelerated device parameters shifts. Two critical requirements should be considered during the execution of the eSM during aging characterization:

1. Accurate device terminal biasing: in order to guarantee a precise voltage biasing avoiding voltage drops due to cables, connectors or metal lines present in array-based IC designs, it is of much importance to use a

Force-&-Sense voltage biasing system. The functionality principles of the Force-&-Sense system will be explained in Section 1.5.3.

2. The time gap between switching from stress to measurement periods should be as short as possible to not loss critical device measurement data at the very beginning of the measurement period. Moreover, it is important that the timing and duration of all stress and measurement phases should be the same for each device and precisely controlled. Otherwise, the data post-processing would get unnecessarily complex and less information would be collected.

Figure 1.14 shows an illustrative example of the eSM methodology executing an accelerated BTI aging test scheme. In the example, four SM cycles are executed sequentially with red and blue colour lines respectively. The initial $I_{DS} - V_{GS}$ curve characteristic utilized to account for the non-degraded device parameters is not shown in Figure 1.14. The eSM technique is conducted, for a single device, with the following the SM repeatable sequence:

1. Stress: device is aged or degraded by setting V_{GS} and/or V_{DS} larger-than-nominal operation voltages, e.g., 2x or 3x the nominal VDD, to accelerate the aging processes testing time. To differentiate between BTI or HCI aging characterization, if $V_{GS}; V_{DS} > 0V$, the accelerated aging is HCI and, if $V_{DS} = 0V$ and $V_{GS} > 0V$, it means that the test is pure BTI. For the eSM presented method, the increase ratio of the stress time has been defined exponential with an initial stress time of 1s (see Figure 1.14).
2. Measurement: in the measurement period, the V_{GS}, V_{DS} biasing conditions are lowered to $V_{GS} \approx V_{th}$ and typically $V_{DS} \leq 100mV$ respectively before changing the device from stress to measurement. The eSM technique starts capturing the $I_{DS} - t$ evolution just before switching the device from the stress to measurement at a high sampling rate. This allows to accurately account for the time gap between the stress and the measurement periods for each device. In addition, a $I_{DS} - V_{GS}$ curve is executed after the $I_{DS} - t$ measurement in every measurement period to further obtain the transistor mobility shift $(\Delta\mu)$ as shown in Figure 1.14 as a blue and green rectangles. During the $I_{DS} - t$ measurement, defect discharges of previously charged defects in the stress period can be observed and, since they are related to $|\Delta V_{th}|$ variations, the equivalent ΔV_{th}-t traces can be also obtained [60, 90].

Typically, when characterizing BTI or HCI aging phenomena, thousands of devices should be tested to get significant statistical results. The eSM test

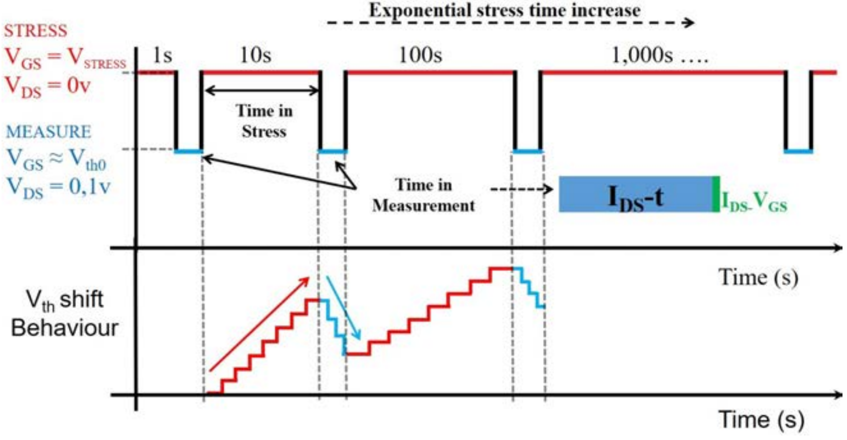


Figure 1.14: Illustrative example of the eSM technique.

patterns are applied equally to all devices involved in the aging test, ensuring the same stress and measurement periods and minimum and equal time gaps between stress and measurement periods. The total test time needed for a serial aging test, i.e., one-DUT-at-a-time, can be calculated from the following equations:

$$T_{SERIAL} = N \cdot (T_{I_{DS}-V_{GS}} + T_{STRESS} + M \cdot T_{MEASUREMENT}) \tag{1.4}$$

with:

$$T_{STRESS} = \left(\sum_{j=1}^{M} K^{J-1} \cdot t_s \right) \tag{1.5}$$

$$T_{MEASUREMENT} = t_{I_{DS}-t} + t_{I_{DS}-V_{GS}} \tag{1.6}$$

where in equation 1.4 N is the number of DUTs and $T_{I_{DS}-V_{GS}}$ is the time to execute the initial $I_{DS} - V_{GS}$ curve. The T_{STRESS} time consists in the total accumulated stress time (see equation 1.5) that depends on the number of SM cycles (M) where the growth rate of the stress time periods is defined by the parameter K, i.e., 10 in this thesis, and t_s is the first stress time period of the initial SM cycle. Finally, the $T_{MEASUREMENT}$ (see equation

1.6) corresponds to the total time required to compute the $I_{DS} - t$ curve and the $I_{DS} - V_{GS}$ during the measurement time.

For instance, let's consider two examples, M5ts1 and M3ts100, to calculate the total test time required for a single device with equation 1.4. The first example (M5ts1), consists in a 5-cycle SM aging test with the following test variables: $T_{I_{DS}-V_{GS}} = 10s$, $M = 5$, $T_{MEASUREMENT} = 100s$, $t_s = 1s$, and $K = 10$. The time required for the aging test of a single DUT would be $T_{SERIAL} \approx 3.23$ hours. For the second example (M3ts100), which consists in a 3-cycle SM aging test where two parameters are changed with respect to the previous example: $M = 3$ and $t_s = 100s$, i.e., the first two SM cycles of example M5ts1 are skipped. Then, T_{SERIAL} is only reduced to 3.17 hours.

In this scenario, when serially performing the same aging tests on many devices, e.g., N = 100 the total test time would rise to approximately 13.4 and 13.2 days for the M5ts1 and the M3ts100 tests respectively, which clearly demonstrates the need for a different alternative when characterizing aging in many samples. To reduce unfeasible aging test times, a parallelization of the stress phases of the eSM technique is used to dramatically speed up the aging tests utilizing array-based IC solutions. In this regard, Section 1.5 will describe a first approach of parallel stress testing methodology named the Parallel-Stress Pipeline-Measurements method and presented in [106].

1.4.2 Defects characterization: The Weighted Time Lag Method

Charge trapping/detrapping provides a highly valuable information of the device physics about device degradation [107, 108]. In this sense, several methods have been proposed in the literature to analyse the defect involved in aging-related and RTN phenomena [32, 33, 109]. However, if the background present in the measurements is relevant enough when compared with the current/voltage steps, the precise defects detection and their parameter extraction can become a tedious task.

The simplest way to analyse charge and discharge of insulator defects consists in represent the data in a histogram to observe the distribution of each point of, for instance, an RTN signal. However, this technique can only be used when the signal fluctuations, resulting from RTN or BTI/HCI aging, are well defined and the background noise is almost negligible. When the current

fluctuations contain a large number of active defects and the background noise is present at the ΔI_{DS} or ΔV_{th} level, more sophisticated methods have to be utilized, like for instance the Time Lag Plot (TLP) [110, 111].

To obtain a TLP, of for instance an RTN signal, the i-th data sample is represented versus the (i+1)-th data sample as shown in Figure 1.15 (a). The main plot diagonal shows populated data regions corresponding to different I_{DS}-t levels while transitions between different levels, as shown in the opposite diagonal of Figure 1.15 (a) as less populated regions. The TLP methodology can be used if the background noise level is low enough, so that the I_{DS}-t levels and/or transitions can be clearly distinguished. However, as shown in Figure 1.15 (b), if the background noise is relevant enough, it can mask capture/emission transitions, resulting impossible to distinguish the different I_{DS} or ΔV_{th} levels.

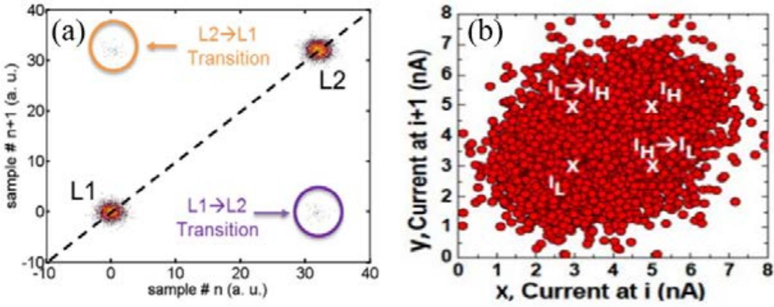


Figure 1.15: Representation of a two level RTN signal TLP where (a) clear discrete levels and transitions can be distinguished [111] and (b) the background noise hides the current levels and transitions [112].

In order to overcome the limitations imposed by the TLP in the presence of relevant background noise, the Weighted Time Lag Plot (WTLP) [112] developed to analyze the charge trapping/detrapping of device defects extends the Time Lag Plot (TLP) procedure being easily implementable and robust when the background noise is large. The WTLP represents the probability that a $(\Delta V_{th-i}, \Delta V_{th-(i+1)})$ data sample in a measured signal, corresponds to one of the I_{DS} or ΔV_{th} levels present in the trace. This probability is calculated by equation 1.7, which corresponds to a normal bivariate distribution function [112] for each measured data sample. Finally, the weighted time lag function defined in equation 1.8, reveals the contribution of all samples in the measured trace, weighted by its distance to the analysed i-th data sample.

$$\phi_i\left(x,y\right)=\frac{1}{2\pi\alpha^2}\cdot exp\left(\frac{-\left[\left(I_i-x\right)^2\right]+\left[\left(I_{i+1}-y\right)^2\right]}{2\alpha^2}\right) \tag{1.7}$$

$$\Phi_i\left(x,y\right)=K\sum_{i=1}^{N-1}\phi_i \tag{1.8}$$

Figure 1.16 (a) and (b) shows the benefits of using the WTLP in an RTN multilevel signal with eight different levels. In this example, if the TLP is used to obtain the I_{DS} levels present, the background makes impossible to locate them as shown in the TLP of Figure 1.16 (a). Nevertheless, as shown in Figure 1.16 (b), where the WTLP have been utilized by means of the Φ function to extract the I_{DS} levels, a higher resolution in the main diagonal of the obtained WTLP have been achieved revealing populated regions across the main diagonal of the WTLP. Each local maximum shown in Figure 1.16 (b), corresponds to an individual I_{DS} level present in the trace.

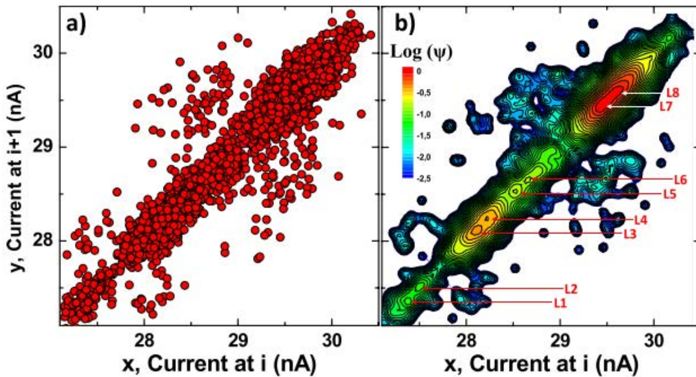


Figure 1.16: Comparison between (a) a Time Lag Plot and (b) Weighted Time Lag Plot obtained from the an experimental RTN signal [112].

In the next section, recently publish array-based IC solutions devised to satisfy the required statistical characterization of transient (RTN) and aging (BTI/HCI) TDV effects reported in the recent literature will be described.

1.5 From wafer-level to array-based IC MOSFET characterization

Originally, variability phenomena characterization utilizing wafer-based systems have been widely used for transistor lifetime and technology reliability prediction as shown in Figure 1.17 (a). This type of device characterization, which implies physical contact on usually one or a few DUTs at a time, results in long serial aging test times that can last months or even years of continuous testing when many devices must be characterized. Moreover, the use of wafer solutions for direct device contact implies that a large area is required to include the necessary pads to access each DUT terminal in the wafer [113].

In order to overcome limitations imposed by on-wafer device testing, integrated circuit chips with thousands of transistors in an array-based configuration have been proposed in recent literature for variability phenomena testing. Figure 1.17 (b) shows an example of the ENDURANCE IC chip designed and fabricated in the core of this thesis to execute statistical device variability characterization. Array-based ICs present two main advantages for the statistical characterization of TDV and TZV: first, a larger number of DUTs can be characterized for a given silicon area and second, precise digital IC control algorithms allow the possibility of utilizing proper parallel characterization techniques that can be used to significantly speed up the statistical characterization of aging degradation phenomena. Moreover, the total area used will be largely reduced, as compared to the wafer-level approach because individual transistor access pads are eliminated from the IC design.

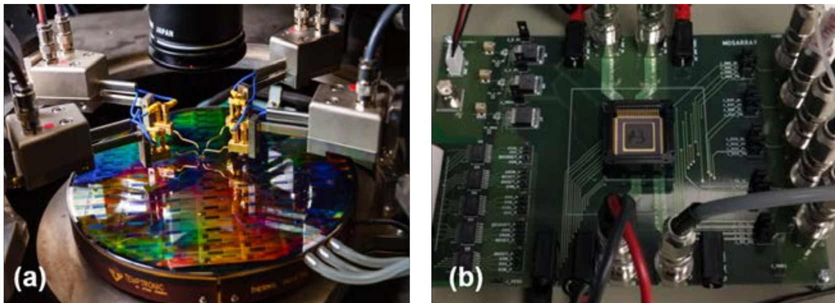


Figure 1.17: Examples of transistor variability testing using a probe station for wafer measurements (a) and utilizing array-based IC chip (b) [114].

In the following subsections, the characterization procedure for proper parallel device testing will be discussed and a full-custom IC measurement framework together with the must-have requirements will be listed to set the basis of transistor characterization when dealing with array-based IC solutions. Moreover, Section 1.5.4 will describe and compare the most relevant and recent IC design structures for variability characterization.

1.5.1 Hardware and software design requirements for array-based IC variability characterization

Despite the advantages of using array-based ICs for device variability characterization, its design is not a simple task when characterizing MOSFET devices under the same conditions as with on-wafer testing frameworks. A set of critical requirements, in terms of IC and measurement system capabilities, should be fulfilled to carry out accurate and trustworthy individual transistor characterization, as listed below:

- Variety of reliability phenomena to be characterized: the IC hardware design should allow the individual characterization of TZV including $I_{DS} - V_{GS}$ and $I_{DS} - V_{DS}$, RTN transient effects, and BTI/HCI aging phenomena all in the same chip.
- Variety of DUTs: the array should include both nMOS and pMOS type transistors, with different geometries to study variability effects in terms of device type and geometry.
- Individual on-chip device access: the array should allow the individual selection of each DUT to separately set its biasing, e.g., to set a DUT on stress or measurement with individual device terminals biasing conditions from the neighbouring devices.
- Accurate device biasing: Force-&-Sense on-chip techniques should be available to precisely apply the required voltages at the DUT terminals to mitigate possible voltage drops. Furthermore, the design should allow high DC current flow during transistor tests, especially for HCI tests.
- Reduction of total characterization time: the array should include the necessary auxiliary circuitry to allow accurately-timed parallel stress testing methods.

- Robustness of selection circuitry: all auxiliary circuitry should be designed to avoid its degradation during the application of DUT stress.
- Accurate current measurement: the array should provide ways to either calibrate or cancel any leakage current that may distort DUT current measurements.

Measurement challenges that an experimental setup should cover to perform full variability phenomena characterization utilizing array-based IC solutions, lead to a set of instrumentation requirements that the measurement system should cover:

- Accurate electrical characterization (which results in the need for cancelling IC leakage currents, applying accurate voltages to the device terminals, minimizing the noise in the measurement lines and ensuring stable biasing) is required to correctly study the dependence of TDV with the bias conditions.
- Accurate thermal characterization over a broad range of temperatures is essential, since the TDV phenomena strongly depend on temperature.
- Characterization times must be affordable.
- Due to the large number of devices and the long characterization times, the control of the test process and data analysis must be fully automated.

1.5.2 Parallel-Stress Pipeline-Measurement method

During the execution of BTI/HCI aging tests utilizing array-based IC chips, parallel testing techniques should be used to drastically reduce the typical prohibitive aging serial testing times. A major constraint for a parallel technique is that, at any given moment, only one DUT should be under measurement so that the collected data only account for the degradation of that specific DUT. By non-violating this rule, DUTs are measured individually while parallelizing stress phases wherever possible. The second aspect concerns the current level when simultaneously stressing a large number of DUTs. This current, typically rising up to mA and is normally directed to a common node (VDD or VSS), and the metal lines carrying it as well as the auxiliary device in the path, e.g., a Transmission Gate (TG), must be sized so that adverse effects like electromigration are avoided.

An interesting parallel technique reported recently in literature is the Parallel-Stress Pipelined-Measurements (PSPM) method [106, 115, 116]. The principle of the PSPM method is to delay the MSM sequence of DUTs with respect to the previous one, resulting in a ”place-and-check” algorithm that partially executes simultaneous (parallel) stress phases and pipeline measurement phases. The main drawback of the PSPM technique is that depending on the duration of the first stress period, the number of DUTs that can be parallelized is limited. Thus, the PSPM technique parallelize a finite number of DUT and repeats the algorithm in series until the non-parallelizable DUTs are tested. Therefore, the PSPM technique becomes inadequate for certain values of the initial stress time and number of SM cycles.

As illustrated in Figure 1.18 (a), the PSPM technique becomes suitable when the stress times, are sufficiently long to measure all involved DUTs before the end of the first stress time. When the initial stress time is short, measurement interferences can occur as shown in Figure 1.18 (b) with red arrows. In the parallel example of Figure 1.18 (b), no more than four DUTs can be tested in parallel and the remaining DUT MSM sequences must be delayed completely increasing the total PSPM time.

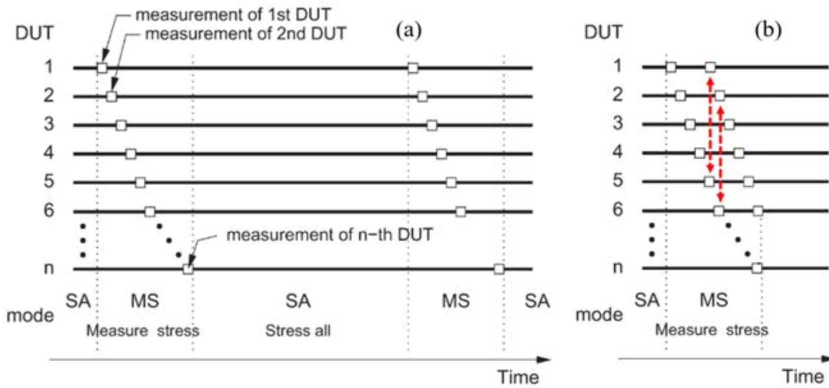


Figure 1.18: Illustrative example of the PSPM technique, where ”SA” and ”MS” stand for stress-all and measure-stress respectively. In (a) the DUTs are measured serially while the rest are stress executed in parallel [117]. In (b) measurement interferences are shown with red arrows reflecting issues related with the PSPM technique when dealing with short initial stress times.

By using the PSPM parallel technique, the maximum number of DUTs (N_{MAX}) that can be parallelized without incurring in a measurement overlap

is given by equation 1.9 for number of DUTs (N). Moreover, the total test time for the PSPM technique (T_{PSPM}) for a number of DUTs below the upper limit imposed by N_{MAX}, can be calculated by equation 1.10 (the initial I_{DS}–V_{GS} characterization time of the fresh sample has been neglected for the total time calculation because it does not significantly influence in the total PSPM time calculation).

$$N_{MAX} = \left\lfloor \frac{t_s \cdot K}{t_{measure}} \right\rfloor + 1 \tag{1.9}$$

where, t_S stands for the initial stress time value, K is the base for the stress time increase, for instance 10 if exponential stress times are utilized, and $t_{measure}$ consists in the fixed measure time period after each stress time.

$$T_{PSPM} = (M \cdot t_{measure}) + \left(\sum_{j=1}^{M} K^{j-1} \cdot t_s \right) + ((N-1) \cdot t_{measure}) \tag{1.10}$$

where N corresponds to the number of devices to be tested, M is the total number of SM cycles, the first bracket accounts for the total time in measurement and the second bracket for the total time in stress (both for one device) and the last bracket computes the final measurement time required when more than one device can be parallelized.

In the works [115] and [116], the described PSPM method is implemented including the necessary delay of the stress/measurement phases in order to avoid any overlap between different devices measurement periods during the aging test. Again, this non-linear delay depends very much on the values of the tuple ($t_{measure}$, t_s, K) as well as on the number of M cycles. The problem with this approach is that, even though some benefit can be gained when compared with the solution in [106], the most typical outcome is that the DUT aging sequence has to be delayed repeating the PSPM tests in series until all DUTs are tested depending on the N_{MAX} boundary condition.

For the previous aging testing examples M5ts1 and M3ts100 reported in Section 1.4.1, equation 1.10 will be used to obtain the parallel test times if the PSPM technique is used instead of serial DUT testing for the two test examples. For the M5ts1 example, the maximum number of devices,

calculated with equation 1.9 can be parallelized is $N_{MAX} = 1$, i.e., parallel stress is not possible because the initial stress time of $t_s = 1s$ is blocking any possible parallelization. However, for the M3ts100 example test $N_{MAX} = 11$ DUTs that can be tested in parallel. Therefore, the PSPM technique becomes inadequate for certain values of the tuple ($t_{measure}$, t_s, K).

For the example M5ts1 above for 100 DUTs, the "place-and-check" algorithm is unable to reduce the test time with respect to the serial method since no device can be tested in parallel resulting in 13,43 days of testing. However, for the example M3ts100 the total test time is reduced to ~ 2.48 days since the larger stress time of the second SM cycle $(t_s \cdot K)$ allows allocation of the measurement of up to 10 other devices during that period of stress.

To overcome issues related to the presented PSPM parallel method, a novel parallelization technique named All DUT Parallel-Stress Pipeline-Measurements, developed in the core of this thesis, will be described in the following chapter. This novel characterization technique has been utilized to significantly reduce the total required testing time for any given ($t_{measure}$, t_s, K) tuple.

1.5.3 Force-&-Sense voltage biasing

Everything used during the device characterization process to connect the instrumentation to the DUT has some innate resistance and capacitance parasitics that should be considered.

To significantly mitigate the effects of cables, connectors and on-chip paths series resistance causing non-desired voltage drops, 4-wire measurement techniques also called Force-&-Sense biasing system can be utilized.

Figure 1.19 shows a Kelvin connection scheme that requires two separate lines for each terminal on the device that should be measured. One pair of lines is used to force current or voltage to the DUT, and the other pair of lines is used to sense the voltage measurement. The key point of this measurement is that the sense lines measuring the voltage ideally are not conducting any current, there is no voltage drop due to cable resistance and the force voltage is raised until the sensed voltage equals the defined voltage [118], [23]. Therefore, the cable resistance effects are significantly mitigated.

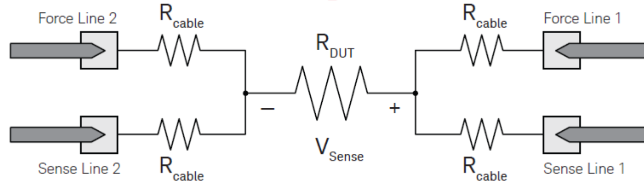


Figure 1.19: A Kelvin (or 4-wire) voltage measurement eliminates the effects of cable resistance by separating out the lines carrying the force current from the lines sensing the voltage [118].

The most common resource for CMOS device measurement is the source/-monitor units (SMU). The SMU can force voltage or current and simultaneously measure voltage and/or current. A noise-free and accurate transistor characterization requires a 4-wire Kelvin measurement thus, SMUs are designed with both force and sense outputs where the force and sense lines are separated to contact the sample for testing, eliminating the effect of cable resistance that induces a voltage drop between the external pad to the device terminal or IR-drop when testing transistors. For instance, Figure 1.20 shows a simplified SMU configuration for a trustworthy device test [118]. In the sense of executing array-based device characterization, Force-&-Sense paths should be included in the IC design to ensure accurate biasing of on-chip devices.

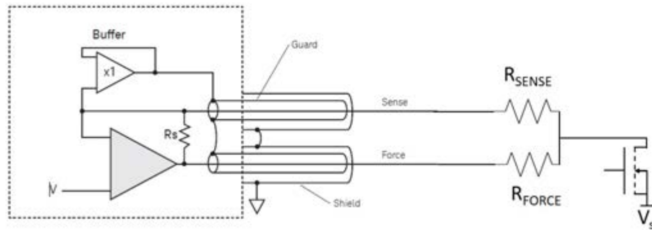


Figure 1.20: Simplified Kelvin SMU output circuit [118].

1.5.4 Previous works in array-based characterization

Regarding the advantages of using array-based IC designs presented in Section 1.5, a few array-based integrated circuits have been reported for the characterization of TZV, BTI, RTN, and/or HCI reliability effects in recent literature. In

this regard, the present section will discuss the advantages and disadvantages of the reported IC works for variability characterization [106,115–117,119–128]. Moreover, Table 1.1 summarizes the set of requirements for each one of the published variability ICs, and shows which of them fulfils the must-have IC design requirements list discussed in Section 1.5.1.

A recent work, which was first used for TZV characterization [119], and later extended to TDV [120,121], describes a very compact array thanks to a simple unit cell design. Figure 1.21 (a) shows the schematic design of the IC chip that allows a high device density (32,000 transistors per chip). The array is also able to perform leakage current suppression. Simplicity and compactness are, however, obtained by sharing the drain terminals of all DUTs in each row and the gate terminals in each column without using intermediate transmission gates (TGs) as shown in Figure 1.21 (b). Therefore, the drain Force-&-Sense paths access the array through a line which is shared by all drain terminals of all transistors in a row. Due to the parasitic resistance of this line, the sensed voltage will be different than the Force voltage applied to each device, introducing voltage drop differences that can affect parameter extraction. On the other hand, without the individual gate and drain TGs, devices in the same column suffer from unwanted stress or, if parallel stress is performed, they are measured at different moments in their recovery phase because only a single device can be measured at the same time to account for individual degradation. Therefore, data need to go through a complex post-processing stage.

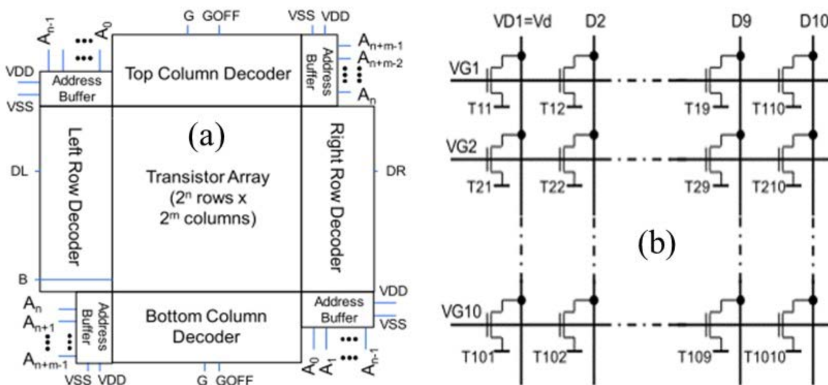


Figure 1.21: Transistor array block distribution where (a) shows the top-level block diagram of a 32k transistor IC and (b) the array of transistors with common gate, drain and source connections without intermediate TGs [119].

When TGs are included to independently access the terminals of each DUT, the number of devices available in the IC array is necessarily reduced due to the area needed to implement the access circuitry, as in [117, 125–128]. The use of TGs to control the biasing of each device also enables accurate timing control during the characterization of aging phenomena. Other arrays also include a large number of DUTs by using simple unit cells designed without TG circuitry [122–124]. The array presented in [122] shows a 65-nm test structure including one million modified SRAM cells, which enable individual measurement of pMOS and nMOS transistors for TZV and NBTI characterization. No Force-&-Sense strategy is implemented, and the gate terminal is shared by all cells in a row, meaning that after stress, devices are serially measured. This is not a major problem in this paper since only the permanent BTI degradation is characterized.

In [124], all DUTs in a row share the gate terminal. Therefore, different portions of the recovery phase will be measured for each DUT, as in [120] and [121]. Furthermore, Force-&-Sense connections are not implemented as in [119], meaning that voltage drops will exist and the precise DUT terminals biasing is somehow unknown. It includes a leakage reduction structure and, unlike the previously reviewed approaches, it considers different DUT sizes.

The chip presented in [123] contains 96x18 cells, each including 48 DUTs and an A/D converter that serially digitizes the current of each DUT. Due to the selected architecture, no Force-&-Sense scheme is necessary, although a calibration is performed for leakage compensation. However, no stress voltages can be applied, and, hence, only TZV and RTN can be measured.

While the array chips proposed in [125] and [128] incorporate on-chip Force-&-Sense paths, those presented in [126] and [117] perform voltage measurement (when a constant current is applied), implying no need for a Force-&-Sense connection, although leakage suppression techniques are used. All four array structures are designed to perform parallel BTI tests, leaving other degradation mechanisms, like HCI, uncharacterised. Similar problems to those described above in [124] are shown in [117], where after a parallel stress, measurements start at different times for different DUTs.

Arrays including TGs usually incorporate two types of transistors in their design: core transistors for characterization and digital circuitry, and I/O transistors for TGs, avoiding the degradation of TG transistors during stress periods. Their operating voltages, when available, have been included in Table 1.1 for a better comparison in the (I/O VDD) row.

Table 1.1: Benchmark of reported array-based IC chips according to the IC must-have list in Section 1.5.1.

Work	[119–121]	[122]	[124]	[123]	[126]	[117]	[125]	[128]	[127]	**[114]**
Tech (nm)	20	65	28	28	180	65	N.A.	28	28	**65**
Devices	32,000	1Mbit 6T SRAM	54,432	¿80.000	3,996	128	1,300	180	5,120	**3,136**
W/L ratios	1	1	6	2	6	8	1	1	1	**8**
CORE V_{DD} I/O V_{DD}	0.9V 1.8V	N.A. 1.8V	1.2V 1.8V	1V N.A.	1.8V 2.7V	1.2V 1.8V	N.A. N.A.	N.A. N.A.	N.A. 2.1V	**1.2V** **3.3**
TZV	YES	YES	YES	YES	YES	YES	YES	YES	YES	**YES**
RTN	YES	NO	YES	YES	NO	NO	NO	NO	NO	**YES**
NBTI	YES	YES	YES	NO	YES	YES	YES	YES	NO	**YES**
PBTI	YES	NO	YES	NO	NO	YES	NO	YES	YES	**YES**
HCI	NO	NO	NO	NO	NO	NO	NO	NO	YES	**YES**
Force & Sense	Limited	NO	YES	NO	NO	NO	YES	YES	NO	**YES**
Temp Range	N.A.	25°C-125°C	N.A.	-173°C-25°C	25°C-135°C	N.A.	N.A.	N.A.	N.A.	**25°C-120°C**
Parallel stress	NO	NO	NO	NO	YES	YES	NO	YES	NO	**YES**
Accurate timing	NO	NO	NO	NO	YES	YES	YES	YES	YES	**YES**
Leakage Current	Partially cancelled	Not cancelled	Partially cancelled	Not cancelled	Partially cancelled	Not cancelled	Not cancelled	Not cancelled	Not cancelled	**Fully cancelled**

While several device arrays have been reported in the literature to characterize BTI phenomena, characterization of HCI phenomena is reported only in [127]. In that work, however, the characterization of aging effects using a parallel technique suffers from unwanted neighbouring device stress and unmeasured recovery time gaps between devices. This IC circuit topology includes TGs only on the gate terminals (10 gate terminals share a TG) while all drain terminals of one row (with 256 DUTs) are connected. Therefore, voltage drops will appear and this will translate into differences in the drain voltage applied to each DUT. This connection scheme implies that when one DUT is being stressed, other DUTs are also being biased and, therefore, degraded, making the data analysis much more complex.

Finally, only [122], [123], and [126] present the characterization of variability phenomena under temperature-controlled conditions. As can be seen in Table 1.1, the unique published array-based IC chip that fulfils all design requirements for a trustworthy variability characterization is the ENDURANCE IC chip [114] being one of the main contributions of this thesis. Design and functionality of our versatile ENDURANCE IC chip will be presented and discussed in detail in Chapter 2 which is capable of accurately and statistically characterizing TZV, BTI, HCI, and RTN in a single array-based IC design.

1.6 Contributions of this thesis

The main goal of this thesis is to increase knowledge and comprehension of the time-zero and time-dependent variability phenomena at the device level in nanometer CMOS commercial technologies. Since unreliability of nanometer MOSFET transistors can affect the intended lifetime and functionality of fabricated VLSI analog and digital ICs, it is essential to perform trustworthy device characterization and provide accurate parameters extraction of the most relevant stochastic sources of variability from a statistical point of view.

In order to obtain statistical data for the evaluation of the damage suffered by individual CMOS transistor under aging mechanisms like BTI and HCI, conventional on-wafer measurements result in unfeasible extremely large laboratory testing times that can grow up to years. To overcome this limitation, in this thesis a novel and versatile array-based IC design, named ENDURANCE chip, recently published in [114, 129], will be described. The versatility of the ENDURANCE IC hardware design permits the execution of time-zero and time-dependent variability tests including trustworthy HCI characteri-

zation. Moreover, the IC integrates eight different transistor geometries and an individual control of the on-chip DUTs for the accurate implementation of parallel stress techniques intended to significantly reduce extremely large aging testing times.

The second major contribution of this thesis consists in the design of a flexible full-custom hardware and software measurements framework that will be utilized to characterize the fabricated ENDURANCE IC chip samples published in [130, 131]. One of the mayor contributions of the measurements system is the creation of a novel parallelization algorithm, designed to significantly reduce the required aging test times by executing in parallel the stress times of a large set of devices while executing individual device measurements to assess the damage suffered by each tested device.

A third contribution of this thesis consists in the obtention of an accurate lifetime prediction based on extensive set of variability measurements over thousands of devices including TZV, BTI and HCI phenomena. This lifetime prediction, based on the analysis of the device's electrical parameters, will give new insights in the comprehension of the different aging variability phenomena at different temperatures. Moreover, this thesis presents different novel procedures to obtain: MOSFET model card parameters, an accurate methodology to accurately capture BTI recovery signals by eliminating RTN and noise contributions, and a novel simulation firework based on CUDA programming for simulating large circuits under aging phenomena in a significant reduced time.

With the understanding of the variability mechanisms of nanometer CMOS technologies, the designed and fabricated ENDURANCE IC chip together with the creation of a full-custom measurements system, this thesis aims to go deeper into the understanding of the most relevant CMOS degradation mechanics affecting devices and circuits designed and fabricated in ultrascaled CMOS technology nodes.

CHAPTER 2

The ENDURANCE array-based Integrated Circuit chip

Due to the stochastic nature of time-zero and time-dependent variability sources, statistical and accurate device characterization must be performed, in the shortest time possible, to account for the effects on device and circuit reliability.

To overcome issues related to conventional on-wafer characterization, as described in Section 1.5 in Chapter 1, several array-based IC structures with thousands of transistors have been presented in the recent literature [114,117,119–128], to characterize TZV and TDV and, in some cases, introduce parallel stress techniques to reduce the overall testing time. Nevertheless, none of the reported works offers an accurate device characterization of TZV, RTN, BTI and HCI TDV in a single IC design and not all of them support parallel stress schemes.

To overcome characterization weaknesses of previous variability testing ICs, this chapter will describe the design of a novel and versatile array-based IC chip, named ENDURANCE, intended for the statistical and accurate characterization of TZV and TDV including HCI phenomenon in individual CMOS transistors. The IC circuit design fulfils all IC chip requirements listed in Section 1.5.1 of Chapter 1. Moreover, its versatile design also permits the implementation of advanced parallel stress algorithms to significantly reduce the time required when characterizing aging phenomena. Parts of the this chapter have already been published in [114, 129, 131].

2.1 ENDURANCE IC design considerations

The ENDURANCE IC, shown in Figure 2.1, consists in a full-custom array-based IC fabricated in a commercial 1.2V 65-nm CMOS technology. The IC integrates all required hardware capabilities for an accurate characterization of the most relevant variability effects like TZV, BTI, HCI and RTN phenomena in a single IC chip.

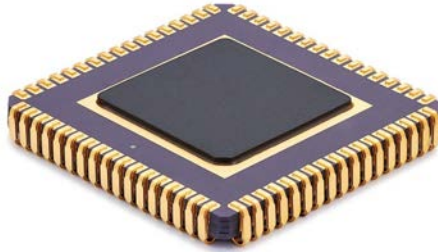


Figure 2.1: The ENDURANCE IC chip encapsulated in a JLCC68 package.

The design of the ENDURANCE IC chip has been focused on accomplish three key design considerations to overcome testing limitations of previous IC designs, as well as traditional on-wafer testing limitations as listed below:

1. During TDV phenomena accelerated aging tests, the digital access circuitry that surrounds the Device Under Test (DUT) can be degraded by the biasing stress conditions generating irreversible damage and destroying the circuit. In order to prevent digital circuitry from unwanted damage in the ENDURANCE chip, each DUT has been designed with an individual digital control circuitry implemented with CORE transistors[1] accompanied with a block of transmission gates fabricated with I/O transistors[2] utilized to bias the DUT terminals while isolating the digital CORE circuitry from external stress voltages.
2. A major concern related to device variability testing utilizing an array-based IC, is the unknown voltage drop that appears between the external chip interface and the transistor DUT terminals. This voltage drop is caused by the internal resistance of the chip metal lines and the

[1]Thin-oxide CMOS transistors operating at 1.2V normally used for digital design.

[2]Thick-oxide CMOS I/O transistors operating up to 3.3V utilized to bias CORE transistors with voltages upto 3.3V

on-resistance of the DUT itself when biased for testing. With the purpose of mitigating voltage drops during DUT characterization, the ENDURANCE IC design has been equipped with internal Force-&-Sense paths to directly sense the voltage at the DUT terminals and adjust the external forced voltage to mitigate possible voltage drops.

3. A dedicated circuit design for variability phenomena testing has been implemented in the ENDURANCE IC which endows enough versatility to carry out parallel stress testing algorithms that significantly reduce aging testing times. To allow this capability, individual DUT Gate and Drain biasing chip paths have been included in the IC design with the aim of separating stress, measure and standby voltages in the chip.

A detailed description of the ENDURANCE IC architecture will be described in detail in the following sections of these chapter.

2.1.1 The IC architecture

The circuitry of the ENDURANCE chip is divided in a set of fundamental building blocks, to facilitate the understanding of the overall IC circuit design. As shown in Figure 2.2, all DUTs of the IC are distributed over two main matrices one for nMOS transistors and another to allocate pMOS transistors named ”NMOS DUT MATRIX” and ”PMOS DUT MATRIX”, respectively. As depicted in Figure 2.2, each DUT matrix has been divided into two submatrices, named ”Left/Right DUT block”, with 56 rows and 14 columns each that allocates 784 DUTs.

The chip allows individual access to the DUTs by means of two row and column twin decoders. To control the selection decoders, individual serial bit input interfaces have been designed consisting of a two-layer D-type flip-flops to store and transmit the row and column DUT addresses to the decoders.

As listed in Table 1.1 in chapter 1, the ENDURANCE chip harbours a total of 3,136 (4x784) 1.2V regular-threshold-voltage MOSFET transistors of eight different geometries including the same number of nMOS and pMOS transistors types. In this regard, Table 2.1, lists the number of DUTs for each geometry. Inside each DUT matrix, devices are distributed as follows: The left DUT blocks contains only DUTs of minimum dimensions, i.e., width W=80nm and length L=60nm, while the right DUT blocks harbours all the geometries.

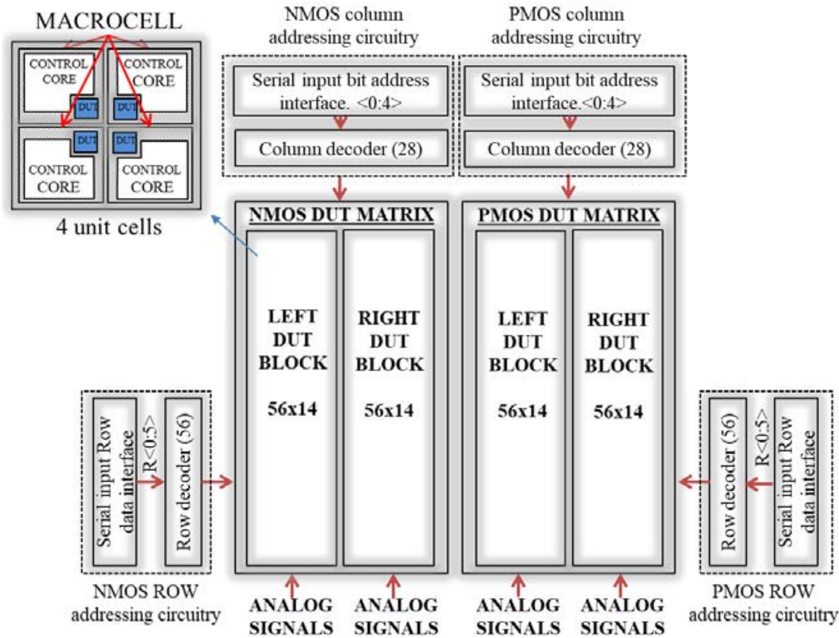


Figure 2.2: ENDURANCE IC fundamental building blocks.

Table 2.1: Total number of unit cells and macrocells in the chip listed as a function of device geometries. Valid for nMOS and pMOS matrices.

Geometry	W/L (nm)	Aspect Ratio	Unit cells	Macrocells
1	80/60	1.33	1376	344
2	200/60	3.33	16	4
3	600/60	10	16	4
4	800/60	13.33	16	4
5	1000/60	16.66	36	9
6	1000/100	10	36	9
7	1000/500	2	36	9
8	1000/1000	1	36	9

Each DUT of the ENDURANCE IC has been fabricated inside a basic and repeatable structure named ”unit cell”. The unit cells are designed with two separated blocks: the ”CONTROL core block”, where the necessary circuitry is embedded for DUT biasing control and the ”DUT block”, where a DUT of a particular geometry is set for electrical testing. Moreover, the inset of Figure

2.2, shows a set of four unit cells which have been vertically and horizontally mirrored and grouped together composing a ”macrocell”.

The circuit design of the unit cell endows high versatility to the ENDURANCE chip because it permits the accomplishment of two of the main design considerations that an array-based IC should cover: first, to isolate the digital control circuitry from the external biasing and, second, to control the biasing conditions of each DUT to execute variability tests for parallel testing purposes. The Layout design of the unit cell will be described in Section 2.1.2 of this chapter.

By means of the control core block, the unit cell can be programmed between four different biasing conditions or ”operation modes”: measurement, stress, standby or off modes as described below:

1. Measurement mode: DUT measurement under nominal voltage conditions, from 0V to 1.2V.
2. Stress mode: BTI and HCI characterization with high biasing voltage, from 1.2V to 3.3V.
3. Standby mode: DUT zero bias: $V_{DS} = V_{GS} = 0V$.
4. Off mode: DUT Drain and Gate terminals disconnection.

The four operation modes described above permit the connection of on-chip devices to different biasing paths inside the ENDURANCE chip. A precise combination and an accurate timing of the described operation modes of the unit cell, permits the execution of user-defined serial and/or parallel stress aging tests as will be detailed in Section 2.2 of this chapter. Furthermore, in order to avoid any electrical damage to the DUTs, during the power-on of the ENDURANCE chip, all DUTs are set to the standby mode by means of a general RESET signal.

To accurately bias Drain and Gate DUT terminals, the unit cells can be programmed to connect those terminals in parallel to one of the three available on-chip biasing signal paths: the measurement, the stress and the standby paths. This parallel connection is done by means of the unit cell transmission gates, five utilized for Drain bias and three for Gate bias. Figure 2.3, shows the schematic diagram of the DUT terminals together with the TG connections at each terminal. Force-&-Sense paths have been designed to access the drain terminal, allowing accurate control of the DUT biasing voltages by minimizing the impact of the IC voltage drop from the IC pads to

the DUT drain. Bulk and Source terminals are short circuited and connected to VDDA for pMOS transistors or VSSA for nMOS ones. For instance, when a DUT is being measured, the drain/gate DUT terminals are connected to the drain MEASURE (F & S) and gate MEASURE (F) path respectively through the corresponding TGs, as shown in Figure 2.3.

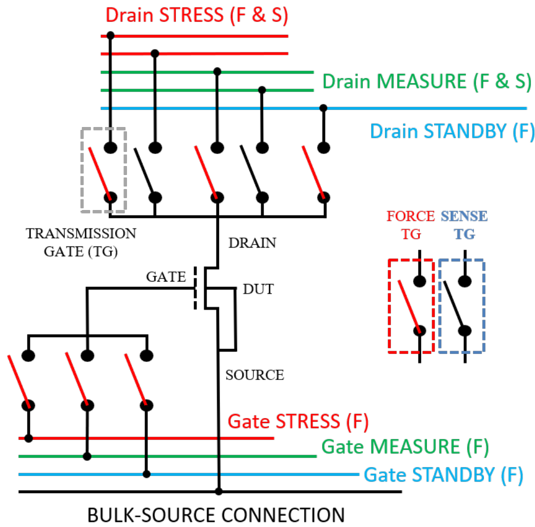


Figure 2.3: Unit cell individual Drain and Gate connections.

Figure 2.4 shows a summarized schematic of the DUT distribution and connections of an IC DUT matrix including all the blocks described in the present section. As previously described, inside the IC, unit cells are distributed across the DUT matrices following an array-based scheme. Unit cells can be individually selected by means of the particular DUT matrix row and column decoders as a part of the selection circuitry. Both row and column input serial bit interfaces are accessed through external digital I/O pads of the ENDURANCE chip.

Different operation modes can be programmed to stablish a parallel connection between the Drain and Gate DUT terminals, by means of the CONTROL core (”CC” in Figure 2.4) and the transmission gates, and the stress, measurement and standby paths of the IC, as shown in Figure 2.4 with red, green and blue biasing paths. Hereafter, the unit cell circuit and layout design will be described in detail.

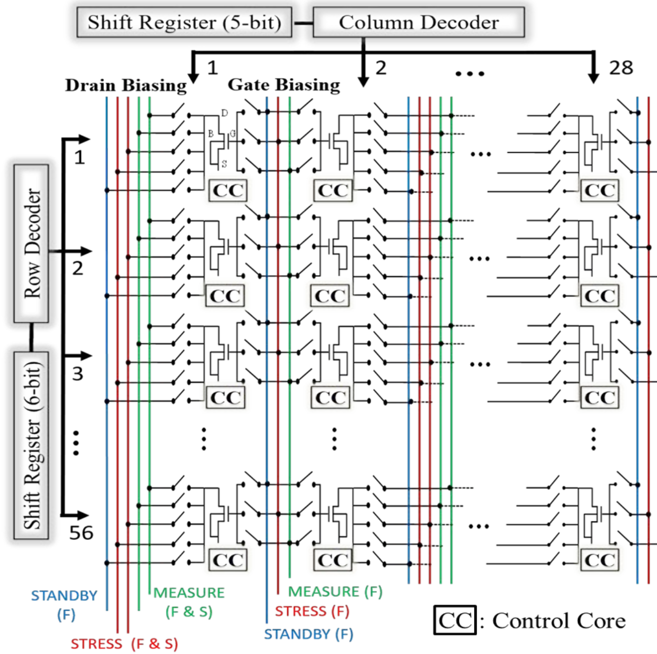


Figure 2.4: Schematic illustration of the unit cell distribution (28 columns and 56 rows) inside an nMOS or pMOS DUT matrix of the ENDURANCE IC chip.

2.1.2 The core: unit cell circuit and layout design

One of the main novelties of the ENDURANCE IC chip is the unit cell circuit design for TZV and TDV measurements. Figure 2.5 shows a schematic representation of the unit cell CONTROL core and DUT blocks together with the external digital control and the DUT biasing signals.

On one hand, the DUT block circuit contains a single nMOS or pMOS transistor and includes independent Drain and Gate connection paths for DUT biasing. On the other hand, the CONTROL core block, from left to right in Figure 2.5, harbours the digital operation mode circuitry, designed to program operation modes to the flip-flop based memories, a level shifter block, utilized to shift the TG control signals from the CORE voltage to the I/O voltage, and the transmission gates (TG) block that connects or disconnects the DUT Drain and Gate terminals from the external biasing.

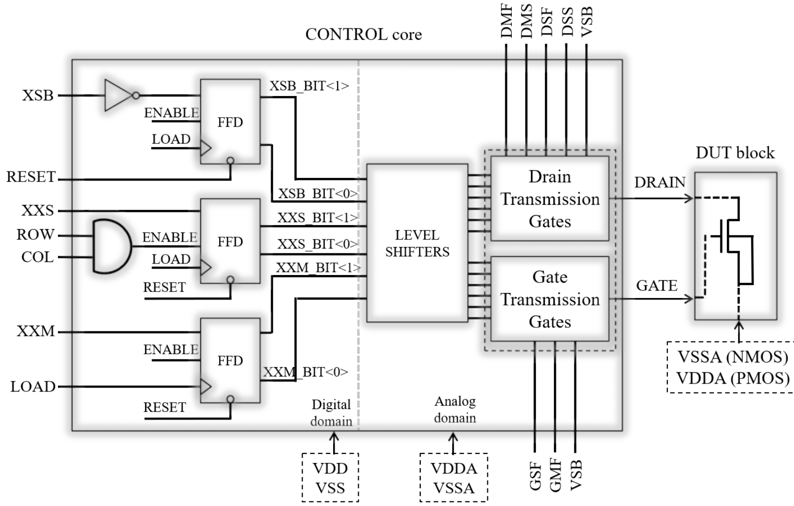


Figure 2.5: Unit cell circuit design composed by the CONTROL core block with digital circuitry, level-shifters and transmission gates, and the DUT block.

The circuit design of the CONTROL core block uses CORE and I/O transistors. The digital circuitry is implemented using only minimum dimension CORE transistors, biased with VDD/VSS (1.2V/0V), while the level-shifter and the TG blocks are only designed with I/O transistors, biased with VDDA/VSSA (2.5V-3.3V/0V), as shown in Figure 2.5, where the voltage domains are separated by a dashed vertical line. The following is a description of the three CONTROL core circuit blocks. Unit cell signal names descriptions can be found in Table 2.2:

The digital control circuitry: this circuit design is based on a set of three single-bit memories, which are implemented with D-type flip-flops, labelled as ”FFD” in Figure 2.5. Each FFD block stores one operation mode bit where the combination of the three bits XSB, XXM and XXS (see Table 2.2) permits to stablish one of the four allowed operation modes to the unit cell. To stablish an operation mode, the FFDs are enabled by means of the ”ENABLE” unit cell signal. The ENABLE signals are the output of an AND gate, where its inputs are the ”ROW” and ”COL” outputs coming from the selection decoders. When the unit cell is enabled, the specific operation mode bits are set to the external IC chip pads and an asynchronous clock pulse has to be sent through the digital ”LOAD” line to store the bits into the FFDs.

Table 2.2: Unit cell signals description.

Signal name	Description
XSB	Standby mode selection bit
XXS	Stress mode selection bit
XXM	Measurement mode selection bit
ROW	Row decoder output
COL	Column decoder output
LOAD	Operation mode establishing asynchronous signal
DMF	Drain measurement force chip path
DMS	Drain measurement sense chip path
DSF	Drain stress force chip path
DSS	Drain stress sense chip path
VSB	Standby voltage chip path
GSF	Gate stress force chip path
GMF	Gate measurement force chip path

Level shifter block: this block serves as a bridge between the CORE and the I/O voltage domains of the unit cell. They perform the conversion of each TG control signal from the VDD/VSS to the VDDA/VSSA voltage domains. Moreover, the level shifters block isolate CORE transistors from unwanted degradation during the application of over-than-nominal voltage DUT biasing.

Transmission gate block: the TG block connects or disconnects the analog IC chips paths for stress, measurement and standby with the DUT Drain and Gate terminals. Drain biasing for Stress and Measurement modes are designed with individual Force-&-Sense chip paths in order to mitigate voltage drops due to internal IC metal lines. TG circuit design will be explained later on in this chapter.

An important step during the IC design phase is the layout design. In this sense, Figure 2.6 shows the compact layout design of the unit cell block. The DUT is located at the bottom right corner where drain, gate and source terminals are labelled in white with ”D”, ”G” and ”S”, respectively. The CONTROL core block is located across the upper side of the layout (section A in Figure 2.6).

Occupying the center part of the layout, the individual transmission gates block, in charge of DUT drain/gate terminals connection for external biasing, is located in section B. Finally, the level shifter block (section C) is located

between Section A and the DUT. To isolate the DUT from any electromagnetic interference from the digital section of the unit cell itself or from neighbour digital lines, the DUT has been surrounded by a well ring (nwell for nMOS and pwell for pMOS DUTs), as shown in Figure 2.7. Moreover, a square metal cover has been inserted on top of the DUT utilizing a top metal connected by vias to the well ring in order to achieve a better DUT isolation from any possible interference as shown in Figure 2.6 with a purple square on top of the DUT area.

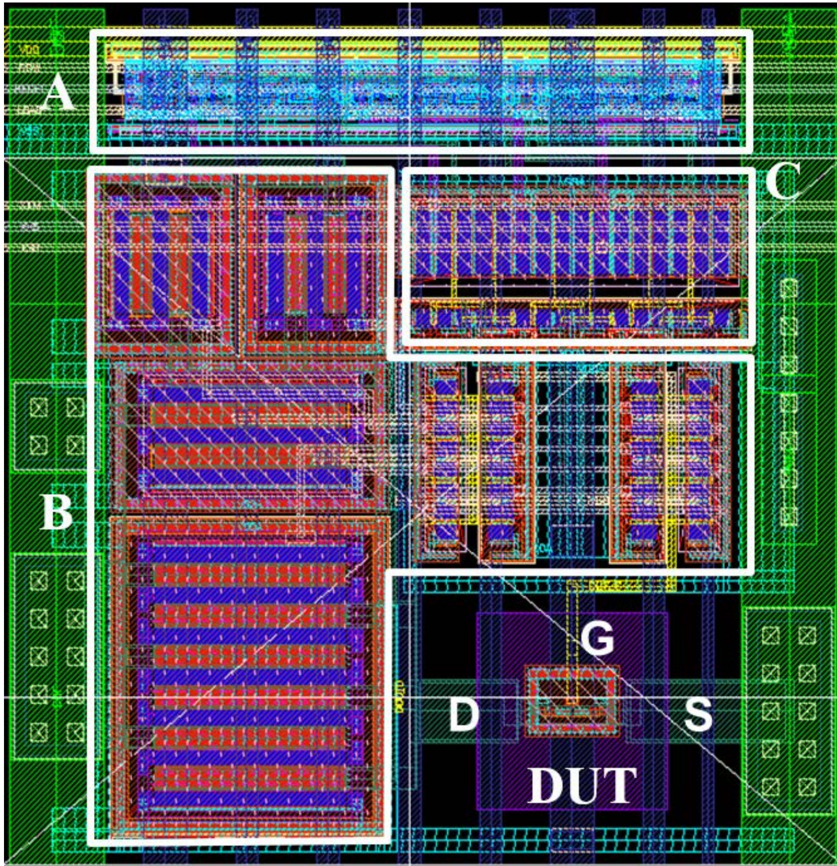


Figure 2.6: Layout design of a unit cell block ($31\mu m$x$30\mu m$) consisting of a dedicated CONTROL core circuitry (section A), the transmission gates block (section B), the level shifter block (section C) and the DUT for electrical test (DUT section).

In order to ensure a precise characterization of the DUTs, two dummy poly gates have been placed close to the DUT transistor of the unit cell as shown in the Figure 2.7, with two red dashed boxes. This technique significantly reduces possible fabrication deviations in the DUT geometry. Source and bulk connections of the DUTs of all unit cells are shared and connected together (VSS for nMOS matrices and VDDA for pMOS matrices) to a metal mesh at the higher metal level to reduce the IR-drop intrinsic to the internal chip metal lines [132, 133].

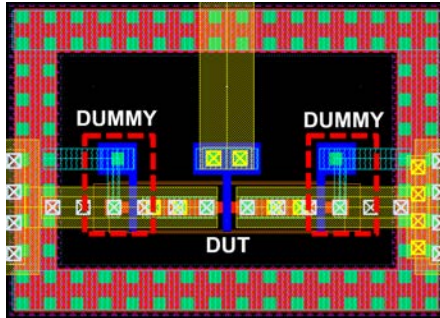


Figure 2.7: Layout design of the DUT block inside the unit cell surrounded by a well perimeter and two side-by-side dummy poly gates.

2.1.3 Unit cell transmission gates block

An improvement of the unit cell circuit design compared to other ICs is the utilization of double TGs to bias DUTs for testing. Figure 2.8 shows an illustration of the unit cell TG block where a set of 5 double TGs have been used for drain stress (DSF/DSS) and measurement (DMF/DMS) purposes with Force-&-Sense paths (see Table 2.2 for signal name description). For the Gate DUT biasing, the measurement (GMF) and stress (GSF) TGs are only provided with a Force path because the current flow is negligible at the gate of a MOS transistor and no appreciable voltage drop will appear. Finally, the standby biasing (VSB) is provided to both drain and gate by means of a force TG at each terminal.

Two important benefits are extracted from utilizing double TGs in the ENDURANCE chip unit cells: first, leakage currents coming from non-selected unit cells are derived away from the measurement channel and, second, double

TG's provide superior performance during circuit operation that is not possible to achieve by using a single transistor switch [134].

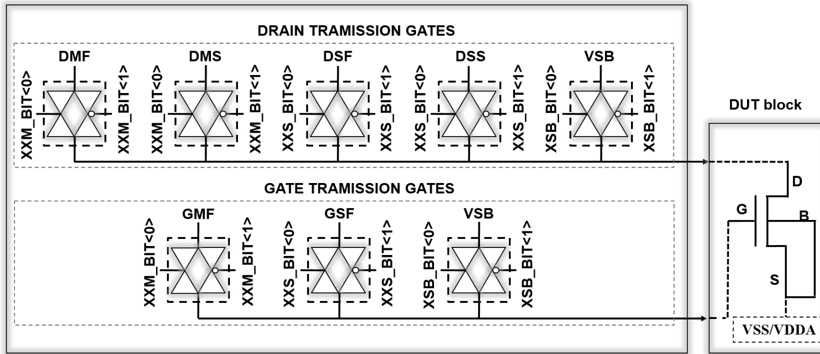


Figure 2.8: Unit cell transmission gate block and DUT block.

In order to mitigate electromigration damage to the TGs due to large current flow and also to reduce the TG layout area, three different designs are used in the unit cell depending on the current flow of each biasing path: the "extra large" TG (TG_XL) used for the DSF biasing path, the large TG ("TG_L") utilized for the DMF biasing path and, finally, the "small" TG ("TG_S") used for the DSS, DMS, VSB, GSF and GMF biasing paths, where the current flow is significantly low compared with stress or measurement TGs.

As listed in Table 2.3, the difference between the three TG designs is the device sizing of the complementary pair of transistors (nMOS and pMOS) of each TG. To reduce the voltage drop from each chip pad to the DUT terminals, all TGs have been designed taking into account the maximum current that will flow through them, and minimizing their resistivity fixing the length at 500nm in all TG designs. Special care has been taken in the sizing of the stress TGs (TG_XL), since high currents are expected in stress mode. Moreover, the width of the metal lines connecting digital and analog paths between all the unit cells TGs have been sized to reduce their resistivity and comply with electromigration design rules. The I/O transistors used in the TG designs were N/P_25OD33_LL type with a lower transconductance than the core MOSFETs.

To avoid ultra-wide layout transistor design, an interdigital layout tech-

nique has been used to save area in the TG layout designs [135]. Figure 2.9 shows the layout design of the TG_XL type where the utilization of the interdigital technique has reduced the area of the pMOS transistor, from $35\mu m \cdot 7\mu m = 245\mu m^2$ to only $10\mu m \cdot 10\mu m = 100\mu m^2$. Moreover, in the TG_XL (see Figure 2.9) type, the pMOS transistor has been designed five times wider than the nMOS one, with the purpose of compensating the low mobility of the pMOS carriers compared to the nMOS transistors. For the TG_S and TG_L types, the dimensions of both nMOS and pMOS transistors have been kept identical.

Table 2.3: Double transmission gates fabricated dimensions.

TG	pMOS		nMOS		Dimensions	Max.
	Width	Length	Width	Length	X, Y (μm)	current
TG_S	1.2 μm	500nm	1.2 μm	500nm	7.4 x 4.5	$500\mu A$
TG_L	3.5 μm	500nm	3.5 μm	500nm	11.17 x 6.7	1.5mA
TG_XL	35 μm	500nm	7 μm	500nm	17.2 x 10	10mA

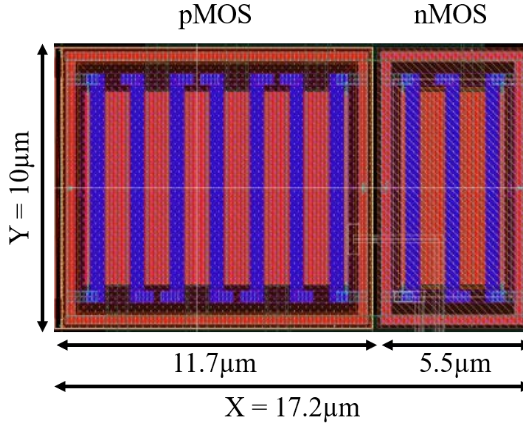


Figure 2.9: Layout design of the transmission gate type TG_XL.

2.1.4 Unit cell level shifter block

The main purpose of the unit cell circuit design consists in allow the execution of TZV and the application of over-than-nominal biasing voltages to the DUT

terminals to accelerate BTI and HCI aging phenomena utilizing the TGs block. In this sense, the activation or deactivation of the TGs is provided by the control circuitry that operates at 1.2V (VDD/VSS). Therefore, all TG control signals have to be shifted from the CORE voltage to the I/O voltage (VDDA/VSSA) by means of three level shifters, one for each operation mode bit.

The three level shifter designs are based in the single supply level shifter type because of the compactness of its layout. The design consists in a set of three cross coupled level shifters improved with the utilization of the Forced PMOS method design technique. This technique consists in breaking down the existing PMOS transistor of the pullup network into two smaller size transistors to significantly reduce the current leakage when the pull up network is activated [136, 137]. The Forced PMOS technique imposes a trade-off between the reduction of current leakage and the delay of the shifted signal when the pullup network is activated.

To evaluate the switching delay of the level shifters in the unit cell, a Cadence transient simulation has been conducted for each type of TG design. The input control signal (labelled as 'IN' in Figure 2.10 (a) has been raised from logic '0' (VSS = 0V) to logic '1' (VDD = 1.2V) imposing rising/falling edge time of 1ns. The shifted OUT and INV_OUT output signals have been connected to each TG type to precisely evaluate the bit setting delay.

Figures 2.10 (b) and (c) show the results obtained from the level-shifter simulations where it can be clearly seen that the larger setting time corresponds to the TG_XL design, as expected. Table 2.4 lists the time delay of the OUT and INV_OUT output signals at the 95% of the VDDA voltage and at the 5% of the VSSA voltage levels. The results show that the implementation of the Forced PMOS technique increases the delay of the control signals when rising from VSS to VDDA voltage domains but as those delays are kept below 10-ns, they can be treated as negligible.

Table 2.4: Level shifter switching transitions for the three TG types.

Transmission Gate type	OUT (ns)	INV OUT (ns)
TG_S $W_{NMOS} = 1.2\mu m$ $W_{PMOS} = 1.2\mu m$	1.40	0.54
TG_L $W_{NMOS} = 3.5\mu m$ $W_{PMOS} = 3.5\mu m$	1.96	0.62
TG_XL $W_{NMOS} = 7\mu m$ $W_{PMOS} = 35\mu m$	3.67	1.69

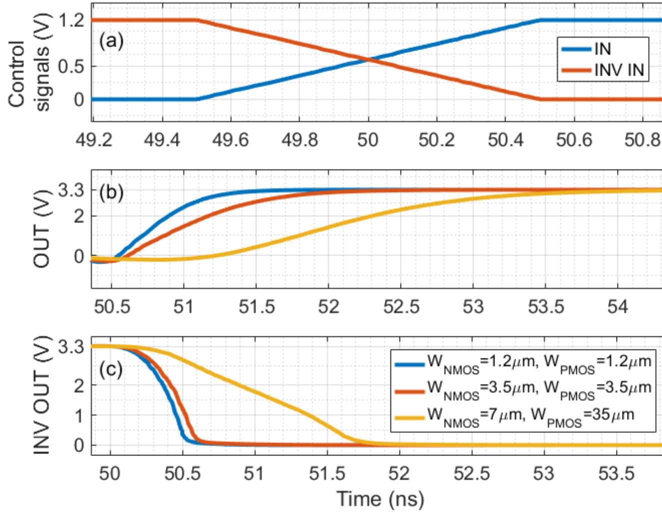


Figure 2.10: Level shifter transient simulation results utilizing an input digital CORE signal with a rising/falling edge of 1ns in (a), the delays of the non-inverting outputs in (b) and the delays associated to the inverting outputs in (c) for the three TG types.

2.2 IC digital control: operation modes

Establishing a precise digital control of the ENDURANCE IC chip is key fundamental to execute trustworthy variability tests. Two main digital functionalities are required by the IC for DUT characterization: the unit cell activation utilizing the row and column selection decoders together with the serial bit input interfaces and the digital procedure for establishing different operation modes to the unit cells.

The design of the ENDURANCE chip incorporates six digital pads, three for the nMOS DUT matrix and three for the pMOS DUT matrix, to externally define one of the four operation modes allowed in the chip for transistor testing. Once a new operation mode is stored into the unit cell memories, the DUT drain and gate terminals are biased for testing. In this scenario, the CONTROL circuitry of each individual unit cell stores and maintains the required operation modes during electrical tests. Thus, depending on the operation mode established, each unit cell DUT will be connected to stress, measurement or standby signal paths as shown in Figure 2.11 independently of

the operation mode established to the neighbour unit cells. The functionality of the four IC operations modes are described hereinafter:

- Measurement mode: the main objective of this mode is to measure I-V characteristics ($I_{DS}-V_{GS}$ and $I_{DS}-V_{DS}$), execute RTN characterization and execute measurement phases after BTI and HCI aging test using the CORE transistor nominal biasing, from 0V to 1.2V. When the measurement mode is established the drain DUT terminal is connected to the DMF and DMS signal paths (see Table 2.2), whereas the gate terminal is connected to the GMF, as shown in Figure 2.11 (a).
- Stress mode: this operation mode has been designed to conduct the stress phases of BTI and HCI aging tests, which consist in the application of an over-voltage biasing, between 1.2V and 3.3V, to drain and gate DUT terminals. When the stress mode is established, the DUT drain terminal is connected to the DSF and DSS signal paths, whereas the gate terminal is connected to the GSF signal path, as shown in Figure 2.11(b).
- Standby mode: this operation mode connects both drain and gate terminals to the VSB signal paths using a single TG for drain and for gate, as shown in Figure 2.11(c). This operation mode has been designed to prevent the DUT from being biased when not selected. For the pMOS transistors, the DUT terminals are biased with VDDA and for the nMOS transistors with VSSA resulting in $|V_{GS}| = |V_{DS}| = 0V$.
- Off mode: This operation mode has been designed as a secure operation mode that opens all TGs of all operation modes leaving the DUT disconnected from the chip biasing paths. The off mode is used to prevent short circuits between the standby mode and other operation modes and to disconnect non-functional DUTs during testing.

All variability test conducted to the on-chip ENDURANCE devices must follow a fixed operation mode protocol with the aim of avoiding any electrical damage to the IC or the unit cells. Possible unwanted stress to the IC paths can damage the chip due to, for instance, a short-circuit between standby and stress voltages if both TGs are activated simultaneously for a long period of time because the operation mode protocol is not followed correctly.

Figure 2.12 shows the allowed operation mode switching protocol executed by the measurement algorithm to ensure a correct unit cell operation during tests. Mode-to-mode critical changes are described below:

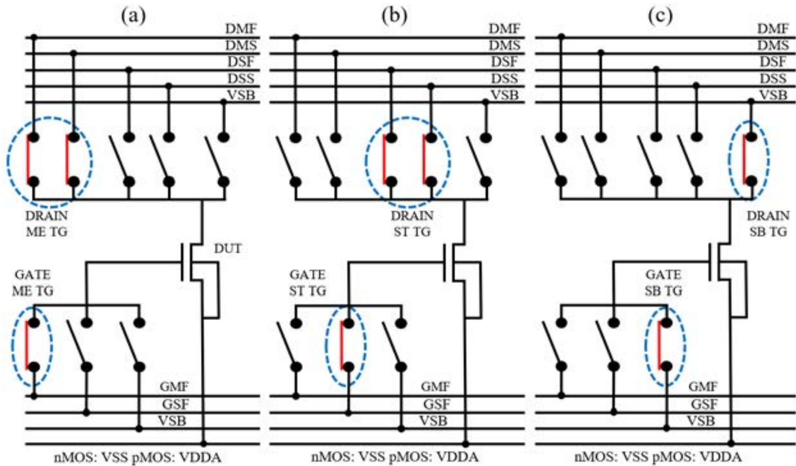


Figure 2.11: Schematic diagram of the TG connections to establish (a) the Measurement mode (ME), (b) Stress mode (ST) and (c) the Standby mode (SB).

- Switching from standby to measurement or stress modes and vice versa: This operation mode switching requires a momentarily Off mode setting to avoid any electrical damage to the internal unit cell transmission gates. A momentarily Off mode ensures that the standby voltage, is not short-circuited with the measurement/stress voltages.
- Switching between stress and measurement modes: switching between these two modes requires an intermediate mode called stress-measurement mode. This special mode ensures that, for a small period of time ($< 10ns$), the measurement and the stress mode internal TGs are connected together. This procedure avoids DUT disconnection that can result in electrical damage or unwanted recovery maintaining the applied voltage at a controlled intermediate level because of the voltage divider imposed by the R_{on} of the TGs.

Figure 2.13 shows an illustrative example of the digital procedure that must be followed to activate a unit cell for operation mode definition using a total of ten digital input signals. As shown in Figure 2.13, the first step before the chip is powered on consists in sending a general reset pulse signal to the chip to switch all flip-flops and decoders to the standby known state. This action will force the selection decoders to activate the non-wired zero output and set all unit cells of both arrays into the standby mode.

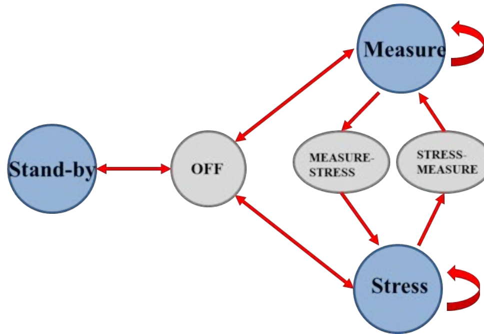


Figure 2.12: Operation mode switching protocol of the ENDURANCE chip.

For activating a unit cell of the ENDURANCE IC, row and column bit addresses are sent bit by bit starting by the MSB through the IROW and ICOL serial data input PADs of the chip. The row and column selection circuitry allow the activation of 56 unit cell rows by a 6-bit word, while the column selection interface manages a total of 28 columns utilizing a 5-bit word address. To store each bit in the serial input interfaces, two asynchronous clock signals need to be sent through the CKC and CKR chip pads. Row and column serial input interfaces are both designed by two-layer interconnected flip-flop lanes. The first lane stores each serially entered bit and, when row and column unit cell bit addresses are entered, the second lane, which is directly connected to the decoders, stores both addresses in memory by means of the asynchronous DECOSET signal. Figure 2.13 shows the unit cell number 5 and 34 activation to switch the unit cells to a different operation mode.

The definition of an operation mode to an active unit cell is conducted by the three digital external chip pads labelled in Figure 2.13 as ”XXM” for the measurement bit, the ”XXS” for the stress bit and the ”XSB” signal for the standby bit. The combination of the three operation mode bits allows for stablishing one of the four operation modes (presented in section 2.1.1) at a time to the unit cell D-type flip-flop memories.

As shown in Figure 2.13, after the activation of the unit cell number 5, the measurement mode is established to the CONTROL core circuitry memories and an $I_{DS} - V_{GS}$ curve is executed to the DUT utilizing the analog measurement paths, i.e., V_{GS}, V_{DS}. For the unit cell number 34, the stress mode is established to the memories and the V_{GS} and V_{DS} biasing voltages are set to 2.5V for the execution of a HCI stress to the DUT utilizing the stress paths.

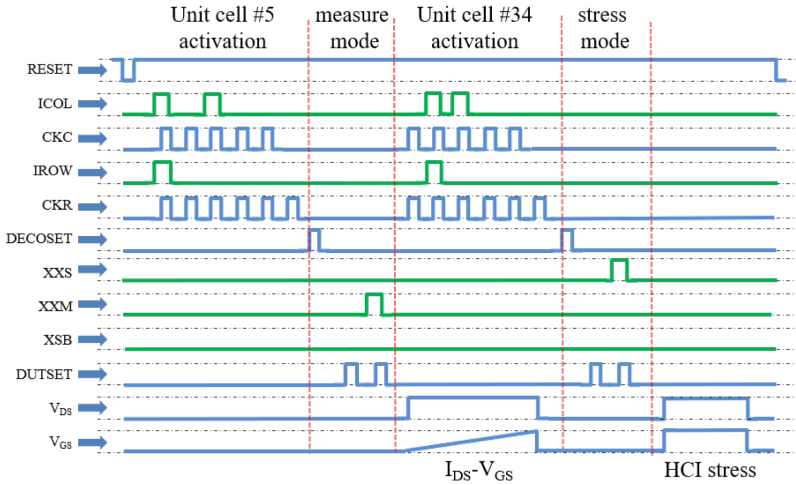


Figure 2.13: Illustrative example of the digital procedure to activate a unit cell for individual operation mode setting and final variability testing.

2.2.1 HCI parallel test example with 3 devices

The present section will illustrate the steps involved in a 3-DUT test example showing the capabilities of the ENDURANCE chip when parallelizing the stress phases whenever possible of an aging test to reduce the total testing time. For this example, an HCI aging test with initial device parameter characterization followed by a single stress-measurement aging cycle will be described utilizing parallel stress phases execution.

The top section of Figure 2.14 describes the graphical distribution of the 3-DUT HCI aging test. The basis of the example consists in parallelizing the stress phases while maintaining individual measurement phases. The non-stressed parameters of each DUT are extracted by means of an initial I_{DS}–V_{GS} (first green block of each DUT in Figure 2.14) curve characterization. During the stress periods of the HCI test, V_{GS} and V_{DS} voltages are both set to 2.5V to accelerate the HCI aging damage. After each stress period during the measurement period (second green square of each DUT in Figure 2.14) an I_{DS}–V_{GS} curve is performed, in this example, with fixed V_{DS}, e.g., 100mV, while the V_{GS} is swept from 10mV to 1.2V.

First, before any electrical test, a secure chip power on is executed. As shown in Figure 2.14, the three biasing chip voltages are set to its required

values: VDD = 1.2V, VDDIO = 2.5V and VDDA = 3.3V. After the chip biasing, all digital circuits of the IC, i.e., shift registers, decoders and the CONTROL core circuitry of all unit cells are at an unknown state. Thus, the RESET signal is lowered to zero and all digital circuits are set to the standby state.

With the aim of saving as much time as possible during the execution of aging tests, the IC should be continuously measuring data from the tested DUTs while the stress phases are executed in parallel. Thus, as shown in Figure 2.14, from the initial $I_{DS} - V_{GS}$ curve characterization of DUT # 5 to the last $I_{DS} - V_{GS}$ curve of DUT #63 after its HCI stress, it can be verified that the IC is continuously measuring $I_{DS} - V_{GS}$ curves, while the stress phases are executed and overlapped as much as possible.

To keep the ENDURANCE IC continuously measuring, it is necessary to include standby periods between certain stress and measurement phases to sustain the integrity of the parallel scheme to reduce the aging test time. Nevertheless, there is an unbreakable condition when executing aging tests changing from stress to measurement phases: a standby period cannot be introduced between the end of a stress phase and the start of a measurement phase for a particular DUT as can be verified in Figure 2.14.

Figure 2.14, also shows when each digital instruction should be sent to the IC to proceed which each step of the complete HCI test. The digital sequence that should be executed at each "DX" stage follows a test priority sequence to ensure a trustworthy device characterization. For instance, in Figure 2.14, in the D6 digital stage DUT # 5 must be switched from stress to measurement while DUT #63 must be switched from measurement to standby. In this scenario, when more than one digital action must be taken the following priorities and obligations should be satisfied:

1. Switching from stress to measurement mode has maximum priority.
2. Switching from measurement to stress mode has low priority.
3. Switching from Standby mode to stress or measurement has low priority.
4. During switching from Stress or Measurement modes to standby mode or vice versa, an intermediate OFF mode should be executed to avoid IC short circuits as defined in the operation mode switching protocol in Figure 2.12. For instance, Figure 2.13 shows the execution of a momentary OFF mode before switching from standby to measurement mode in DUT #5 and from standby to stress mode in DUT #34.

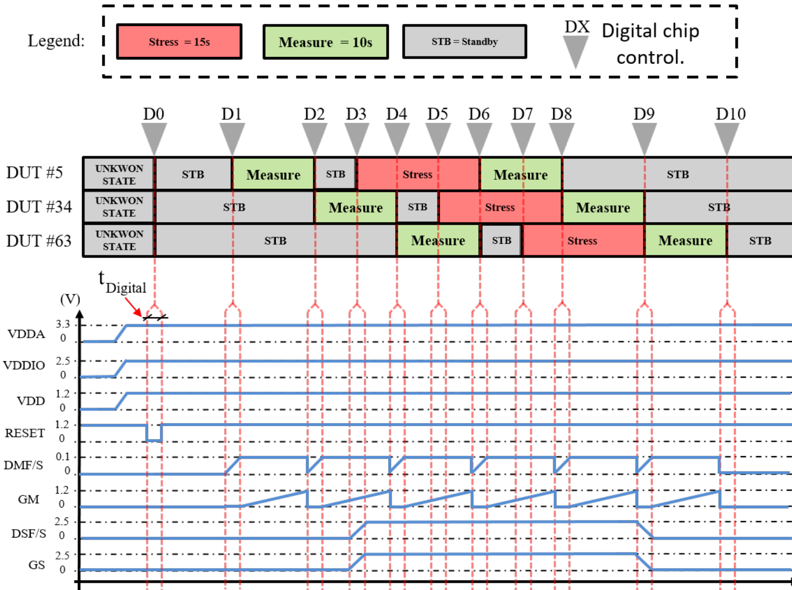


Figure 2.14: Timing sequence of the 3-DUT HCI test example.

A precise description of the novel parallel algorithm shown in Figure 2.14, named All DUT parallel stress pipeline measurements ADPSPM, will be described in detail in the following chapter of these thesis.

2.3 Conclusions

In this chapter, the circuit design of the novel and versatile ENDURANCE IC chip for the massive characterization CMOS transistors has been presented. The chip contains sufficient test devices (nMOS and pMOS) of different sizes to get statistically significant results for a wide range of variability test including TZV, RTN, BTI and HCI aging that can be characterized using a unique IC.

Compared to other existing array-based ICs, the ENDURANCE chip incorporates an accurate circuit design that ensures the ability to perform trustworthy device level variability characterization. The utilization of on-chip Force-&-Sense voltage biasing paths guarantees that, during variability characterization, IR-drops are mitigated and all defined voltages are correctly

applied to the device terminals. The IC I/O TGs allow applying stress voltages to the DUTs up to 3.3V during aging tests without significant degradation of the access circuitry. The DUT circuitry added for this purpose, however, limits the number of DUTs that can be included.

The digital circuit design incorporated in to each DUT together with the dedicated TG allows individually connecting each DUT embedded in the unit cell in parallel to the measurement, stress or standby paths of the chip. This functionality allows the execution of stress phases of aging tests in parallel, significantly reducing the test time. Moreover, the IC design ensures accurate synchronization between measurement (with leakage current cancellation) and stress phases. The completely asynchronous unit cell selection circuitry increases the IC program flexibility because row and column address bits can be introduced in parallel providing faster unit cell activation. Moreover, a new unit cell selection address can be pre-programmed in the shift registers saving valuable time.

Special care has been taken in the layout design of the unit cell block. The interdigital technique utilized in the design of the TGs allows a significant area reduction of the total size of the unit cell. The layout design of the DUT itself, includes side-by-side dummy transistors to minimize DUT fabrication geometry mismatch. Another advantage of the ENDURANACE chip is the possibility to study TZV and TDV related variability phenomena in terms of DUT size, distance and orientation, owing to the novel array distribution in macrocell blocks.

CHAPTER 3

A flexible framework for the characterization of TZV and TDV phenomena in array-based ICs

To overcome issues related to on-wafer transistor testing, as described in Section 1.5 in Chapter 1, array-based IC solutions allow the statistical characterization of transistors but brings new characterization challenges, which are critical not only for the design of the IC itself but also for the way in which the on-chip devices should be tested. Therefore, efficient characterization frameworks must be developed to provide a hardware structure to harbour the array-based IC chips together with a fully automated software tool to generate and precisely execute all required control instructions during transistors testing. In this scenario, the present chapter will describe the full-custom laboratory instrumentation framework together with a fully-automated software tool intended for the characterization of CMOS devices in array-based IC solutions. Parts of these chapter have been published in [130, 131].

This chapter is divided in four sections. In the first section, the novel parallel stress testing technique for BTI and HCI aging phenomena characterization will be described. The second section will present the full custom characterization software, named TARS, to perform statistical characterization of TZV and TDV phenomena automatically. The TARS software has been designed using Matlab® scripting language, with a set of easy-to-use graphical user interfaces (GUI) for the creation and execution of automated tests. The third section will explain in detail the developed characterization hardware that harbours all necessary blocks to characterize the on-chip ENDURANCE transistors. Finally, the fourth section will cover the digital and analog ENDURANCE IC software functionality verification.

3.1 The All DUT Parallel Stress Pipeline Measurements (ADPSPM) methodology

Regarding prohibitive testing times when dealing with conventional BTI/HCI testing techniques, parallel stress techniques are used to dramatically reduce the overall test times when thousands of MOSFET devices are involved. In this regard, the stress phases can be parallelized but only one DUT can be measured at a time to assess the degradation of each device individually. The PSPM technique, described in Section 1.5.2 in Chapter 1, deals with this constraint and, as it has been demonstrated, this methodology cannot apply a good parallel scheme when dealing with an MSM sequence with short initial stress times.

In order to avoid the limitation imposed by the PSPM method, a novel parallel algorithm has been developed in this thesis in order to execute BTI and HCI aging tests in the shortest possible time accommodating hundreds of devices in a single test. This novel methodology, named All DUT Parallel Stress Pipeline Measurement (ADPSPM) uses standby periods[1] that can be introduced between certain measurement and stress phases to make the necessary room to accommodate any number of DUTs in the parallel scheme. These intermediate standby periods will be considered in the data analysis (as an additional recovery time) for the evaluation of the damage suffered by the tested DUTs under BTI or HCI accelerated aging test. For instance, Figure 3.1 shows an example of an arbitrary 5-DUT aging test with four eSM cycles parallelized by utilizing the novel ADPSPM algorithm.

With the aim of achieving a precise SM parallel test, the ADPSPM algorithm treats and analyses each cycle SM cycle individually. In the first cycle (C1 in orange in Figure 3.1), the SM pattern of each DUT is delayed ("D" times), like in a regular PSPM approach, in order to ensure that all the measurement phases in all DUTs are pipelined and no measurement overlap occurs. Once all the SM patterns of cycle C1 have been temporarily allocated, the algorithm starts the parallel distribution of the next cycle. The critical condition is that the last measurement phase of cycle C1 (DUT#5 in Figure 3.1) cannot overlap with the first measurement phase of cycle C2 (DUT#1). In order to guarantee this unbreakable condition, the algorithm computes the total SM time needed for DUT#1 in cycle C2 (time B), and compares it to

[1] A standby period is a time frame where a DUT is biased with $V_{GS} = V_{DS} = 0V$, which means that the DUT is not being measured not stressed.

the time required for measurement of the remaining DUTs in cycle C1 (time A). If the difference is negative ($B < A$), the algorithm inserts the necessary standby time period (SB) to avoid any measurement overlap, shown in Figure 3.1 as green periods.

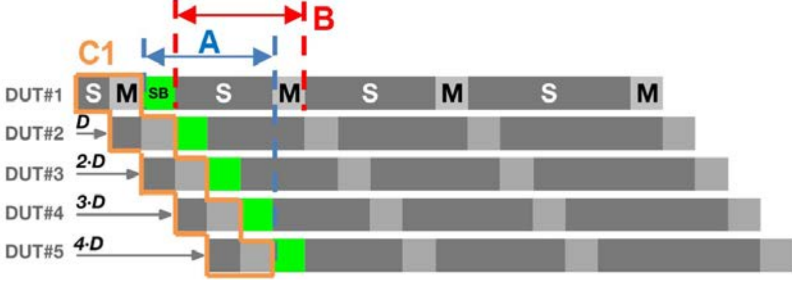


Figure 3.1: Example of the ADPSPM technique with 5 DUTs during a 4-cycle SM test. C1 illustrates the first SM pattern. A denotes the time reserved for DUT measurements after the end of the DUT#1 SM cycle and B is the required time for the SM pattern execution of DUT#1 in the next cycle.

The ADPSPM parallel methodology can accommodate any number of DUTs avoiding any "place-and-check" DUTs allocation. The total test time of the ADPSPM algorithm can be calculated with the equation 3.1:

$$T_{ADPSPM} = t_s + (M \cdot N \cdot t_{measurement}) + t_{SO} \tag{3.1}$$

$$t_{SO} = \sum_{j=2}^{M} max \left[\left(K^{j-1} \cdot t_s - t_{measure} \cdot (N-1) \right), 0 \right] \tag{3.2}$$

In equation 3.1, the t_s time corresponds to the initial stress time defined to all DUTs in the eSM sequence. The first bracket computes the total time where the system is continuously measuring devices, computed with the number of SM cycles (M), the number of DUTs to be tested (N) and the fixed measurement time ($t_{measurement}$) for all DUTs. Finally, the "Stress only" additional time period (t_{SO}) stands for the total amount of time where the system is only stressing devices in an i-th eSM cycle. This is because before the end of the stress period in the i-th eSM cycle, all DUT measurements of the (i-th)-1 cycle have been concluded. In this sense, in equation 3.2, K

corresponds to the stress time increase base, e.g., K = 10 for exponential increase, and j defines eSM i-th cycle. This situation can occur when large stress times, e.g., 10,000s or 100,000s, combined with a low number of DUTs in the test are defined to the ADPSPM algorithm for its parallelization.

Figure 3.2 shows an explanatory eSM test parallelized with the ADPSPM methodology for an aging test with M = 5 cycles, K = 10, $t_{measurement}$ = 100s, t_s = 1s, and N = 5 DUTs. In the 5-th eSM cycle, each DUT is stressed for 10,000s while the 5 DUTs in the 4-th cycle are measured in $N \cdot t_{measurement} = 5 \cdot 100s = 500s$, a significant reduced time when compared with the upcoming stress period of 10,000s. In this situation, after the end of the last DUT measurement in the 4-th cycle (T_1 in Figure 3.2) the system will be only stressing the 5 DUTs at the 5-th cycle for 10,000s - 500s ≈ 2.7 hours (using equation 3.2 with j=5) until the first DUT stress period is concluded (see T_2 in Figure 3.2). On the contrary, if the same test is executed increasing the number of DUTs, for instance N = 100, the t_{SO} computed time with equation 3.2 at the 5-th cycle ($j = 5$) is reduced to only 100s, meaning that the number of DUTs in the test influences the total amount of t_{SO} of the entire parallel scheme.

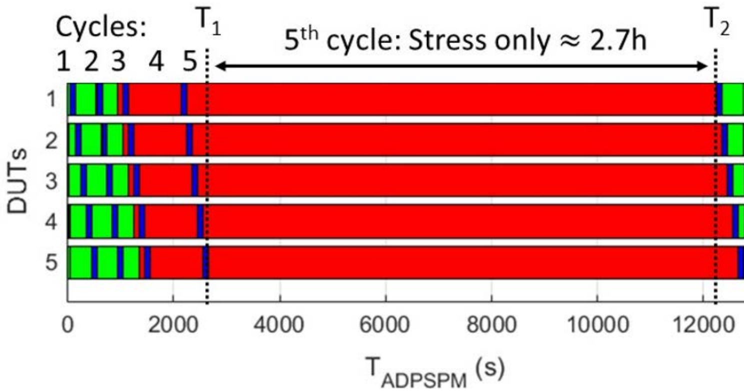


Figure 3.2: Example of the time in "Stress Only" in a ADPSPM test with 5-th eSM cycles due to a low number of parallelized DUTs combined with long stress times. In the Figure, the 4-th cycle shows a lower t_{SO} time.

In order to illustrate the time reduction achieved by the novel ADPSPM technique, Table 3.1 shows a comparison in terms of total aging test time between SERIAL, PSPM and the presented ADPSPM algorithm for the two

aging test examples, M5ts1 (M = 5, $t_s = 1s, t_{MEASURE=100s}$) and M3ts100 (M = 3, $t_s = 100s, t_{MEASURE=100s}$), described in section 1.4.1 of Chapter 1 for different number of DUTs, i.e., 1, 100, 300, 500, 700.

Table 3.1: SERIAL, PSPM and ADPSPM methodologies test time comparison

Number of DUTs	SERIAL	PSMP	ADPSPM	SERIAL	PSMP	ADPSPM
	Test: M5ts1			Test: M3ts100		
1	3.22h			3.16h		
100	13.4d	13.4d	13.9h	13.2d	1.45d	8.4h
300	40.3d	40.3d	41.7h	39.5d	4.05d	1.04d
500	67.2d	67.2d	2.90d	65.9d	6.65d	1.74d
700	94.1d	94.1d	4.05d	92.3d	9.26d	2.4d

As shown in Table 3.1, it has been clearly demonstrated how the serial device test takes the longest characterization time. Moreover, the PSPM technique cannot go below the serial testing time if the initial stress time if set to 1s for the M5ts1 test definition. This happens because the PSPM algorithm cannot parallelize the SM sequence because measurement conflicts appear between different devices and cannot be avoided. Thus, the SM sequences have to be delayed increasing the test time.

For the M3ts100 test, where the initial stress time is increased to 100s, the PSPM shows a good aging test time reduction executing the tests approximately 10 times faster than the serial procedure for all DUT number cases. In this case, the PSPM is able to parallelize more than one device without incurring in measurement conflicts but, still with an initial large stress time, the PSPM technique is not able to significantly reduce the total aging test time. On the contrary, the total test time reduction achieved by the ADPSPM technique is significantly larger when compared with the SERIAL and the PSPM methods. For instance, for the test involving 700 DUTs, the ADPSPM technique executes the M5ts1 test 23 times faster than the SERIAL and the PSPM techniques or, equivalently, the ASPSPM technique achieves a test time reduction of 95% (see Table 3.1).

For the case of the M3ts100 test example, the ADPSPM technique executes the aging test ≈ 4 times faster than the PSPM algorithm or, equivalently, achieves a test time reduction up to 74%. Typical BTI or HCI aging tests intended to accelerate aging phenomena are based in the execution of MSM

patterns similar to the M5ts1s test example or even with starting with lower stress times [50, 60, 127, 138].

3.1.1 ADPSPM: a time complexity study

To illustrate the advantages of the ADPSPM technique in terms of time saving, the present section will describe a study of the required time to characterize DUTs comparing the SERIAL, PSPM and ADPSPM algorithms against an IDEAL test time frontier. For the comparison, Figure 3.3 shows the characterization time per DUT as a function of the number of DUTs for three different aging tests: a 5-cycle eSM test with initial stress time of 1s ($t_s = 1s$), a 4-cycle eSM test with $t_s = 10s$ and a 3-cycle eSM test with $t_s = 100s$ shown in Figure 3.3 (a), (b) and (c) respectively. Moreover, to compute the test time per device, the measurement time of all measurement periods for all DUTs is set to a 100s $t_{measurement} = 100s$.

As shown in Figure 3.3 (a), (b) and (c), the blue diamonds represent the time required for the SERIAL test in which the devices are characterized one at a time, by applying the full sequence of eSM cycles. The test of the following DUT does not start until all eSM cycles of the previous device have finished. If we consider the place-and check PSPM technique for the conditions shown in Figure 3.3 (a) in red squares, no parallel stress can be applied because the initial stress time of 1s does not allow any parallelization of the SM cycles of different devices thus, only one single device can be tested at a time overlapping with the SERIAL testing time (see inset in Figure 3.3 (a)). This means that no time improvement when compared with SERIAL testing can be achieved.

On the contrary, the ADPSPM technique clearly shows a fast decrease of the test time per device when increasing the number of devices, as shown with the green circles. In Figure 3.3 (b), the initial stress time is set to 10s allowing the PSPM technique to parallelize only 2 devices, whereas in Figure 3.3 (c) the initial stress time is set to 100s and up to 11 devices can be stressed in parallel with the PSPM technique. Correspondingly, the test time per device is reduced to one tenth of the SERIAL procedure, but this 2-device or 10-device parallel scheme must be repeated until all devices are measured.

Also in these cases, the ADPSPM technique reveals a significant test time reduction compared to the SERIAL and PSPM techniques, by parallelizing

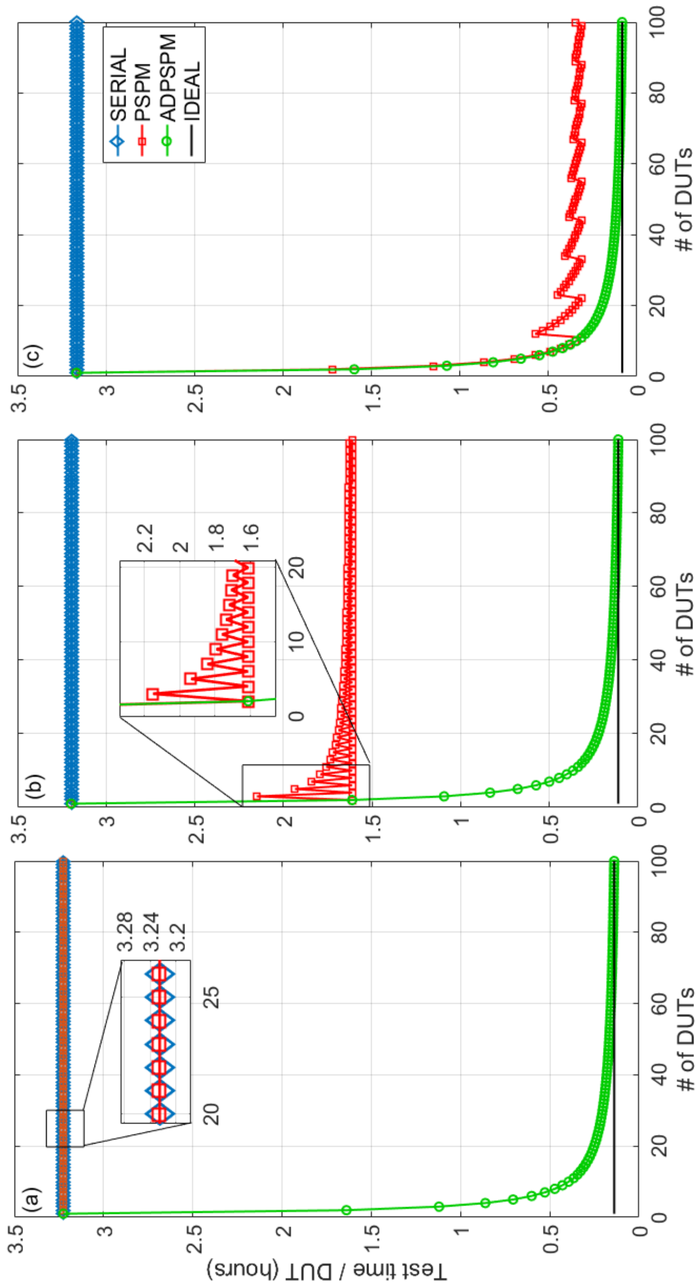


Figure 3.3: Total test time per DUT as a function of the number of devices for SERIAL, PSPM and ADPSPM techniques considering a measurement time of 100s and an exponential increase of stress time of the SM cycles (base 10) with (a) 5-cycle SM test starting with a stress time of 1s, (b) 4-cycle SM test starting with a stress time of 10s and (c) 3-cycle SM test starting with a stress time of 100s.

the stress phases whenever possible and maintaining the measurement channel occupied the maximum time (see Figure 3.1). The IDEAL time is indicated as a black line in Figure 3.3 and represents an unreachable limit of the test time per device where the stress time of the necessary parallelization scheme are ignored. It can be checked in Figure 3.3 (a), (b) and (c) that in all test situations, the ADPSPM technique tends asymptotically to the IDEAL time when the number of DUTs increase.

3.2 TARS software architecture for MOSFET characterization

During transistor variability characterization, it is very important to have a precise control of every action taken by the measurement system such as the particular biasing conditions at a given time during stress or measurement periods. Moreover, maintaining a precise synchronization between sending digital instructions to the IC chip and executing the required characterization line, for instance, $I_{DS} - V_{DS}$ curves, is also critical. In this scenario, all hardware components of the measurement framework must be precisely controlled by a set of GPIB[2] [139] and USB instructions. These important steps could be implemented by the user itself by hand sending each instruction with a perfect synchronization between all the setup equipment. Nevertheless, this manual synchronization would become an unfeasible task when thousands of commands have to be sent to the laboratory equipment in a specific sequence maintaining a perfect synchronization.

For illustration, let us consider a simple $I_{DS} - V_{GS}$ curve characterization for a chip containing 100 devices. Assuming that 100 sampling points are taken for each curve, approximately 10,000 GPIB instrumentation commands are needed. In the case of a typical BTI/HCI degradation test, not only the number of commands will be even larger but their scheduling will be much more complex. Therefore, the need for automating the generation of instructions for the control of the hardware setup becomes necessary. In this scenario, the TARS software generates all necessary GPIB and digital control instructions. The TARS software features several user-friendly GUIs that facilitate the definition of automated measurement tests that will be described in the following subsections of these chapter. The architecture of the TARS

[2] The General Purpose Interface Bus is a digital communications interface bus with specification IEEE 488.2

software follows a modular approach with the following three main modules:

1. Test plan generation module: this module receives the information about the user-defined tests and generates a script[3] containing a specific "test plan". A "test plan" file consists of a proper sequence of the electrical tests to be applied to each DUT and the range of the test conditions.
2. Elementary instruction generation module: this module transforms each command of the test plan into the set of elementary actions that implement such command.
3. Test execution module: this module reads each elementary instruction and generates and sends the corresponding GPIB commands to the instrumentation and the IC digital control instructions.

The three presented modules, that constitute the architecture of the TARS software, are based on the automatic generation of test files, for the configuration of the user-defined test using the "Test plan generation" and the "Elementary instruction generation" modules and also to store the measured data from the tested devices. The structure of the generated files follows a fixed column distribution containing all the required data to execute tests and store measured data as shown in Table 3.2. The main column data distribution of all generated files shares exactly the same 5 initial columns with the test time, DUT number, MOSFET type, operation mode (OM) and, variability code (VC). Moreover, the sixth column is expanded into a set of columns depending on the type of test to be executed. The column distribution of the files generated by the TARS modules will be explained in the following sections.

3.2.1 Test plan generation module

The main objective of the test plan generation module consists in the automatic generation of a text file that contains a test plan where all the required information of the variability test to be executed is specified in individual test lines. The test plan is generated from the user-specified test options entered by means of the test file generation GUI that will be described Section 3.4. The structure of the test lines contains the test variables distributed in columns following the structure presented in Table 3.2. Depending of the test

[3]Files where programs are written utilizing high-level general-purpose languages for a special run-time environment to automate the execution of tasks.

to be executed (see variability test codes in column #5 of Table 3.2), the test lines can have 7 or 10 columns.

Table 3.2: Column distribution.

1	2	3	4	5	6
Test time	DUT number	MOSFET type	OM	VC: Variability code	Biasing conditions
Seconds	Integer	N: nMOS P: pMOS	SB ME ST	00: standby 11: stress 22: BTI measure 23: HCI measure 33: $I_{DS} - V_{GS}$ 44: $I_{DS} - V_{DS}$ 20: RTN measure	Initial Final Step V_{GS} V_{DS}

For the automatic generation of the test plan file, the algorithm executes the flow diagram of Figure 3.4, which is divided in three different branches: the I-V measurement branch, dedicated to describe the generation of $I_{DS} - V_{GS}$ and $I_{DS} - V_{DS}$ curves, the RTN test lines generation branch, and the BTI/HCI aging test lines branch generation, which applies the ADPSPM technique to all DUTs following the test line instruction described hereinafter:

(i) $I_{DS} - V_{GS}$ curve test line column variables: Table 3.3, lists the 9 column variables, where columns 6 to 9 allocate the required variables to execute an $I_{DS} - V_{GS}$ curve. Table 3.3 shows two examples of $I_{DS} - V_{GS}$ curve definition for nMOS and pMOS DUTs, where the VC is 33 as defined in Table 3.2.

Table 3.3: $I_{DS} - V_{GS}$ test line column variables definition.

1	2	3	4	5	6	7	8	9
Test time	DUT#	MOSFET type	OM	VC	V_{DS}	Initial V_{GS}	Final V_{GS}	Step V_{GS}
5	15	N/P	ME	33	0.1	0.01	1.2	0.01

(ii) I_{DS}–V_{DS} test line column variables: Table 3.4, shows the 10 column variables where columns 6 to 10 allocate the required variables to execute an I_{DS}–V_{DS} curve, where the VC is 44 as defined in Table 3.2.

Table 3.4: $I_{DS} - V_{DS}$ test line column variables definition.

1	2	3	4	5	6	7	8	9	10
Test time	DUT#	MOSFET type	OM	VC	Initial V_{DS}	Final V_{DS}	Initial V_{GS}	Final V_{GS}	Step V_{GS}
5	15	N/P	ME	44	0.01	1.2	0.01	1.2	0.01

(iii) RTN and BTI/HCI aging test lines definition: the test plan generation algorithm uses the same 8-column variables. Table 3.5 shows 5 different example test lines where, line #1 defines the variables for the RTN test with VC code 20; line #2 and #3 define a SM test for a BTI execution with VC 11 and 22 respectively; lines #4 and #5 define the SM test lines for an HCI test with VC 11 and 23 respectively.

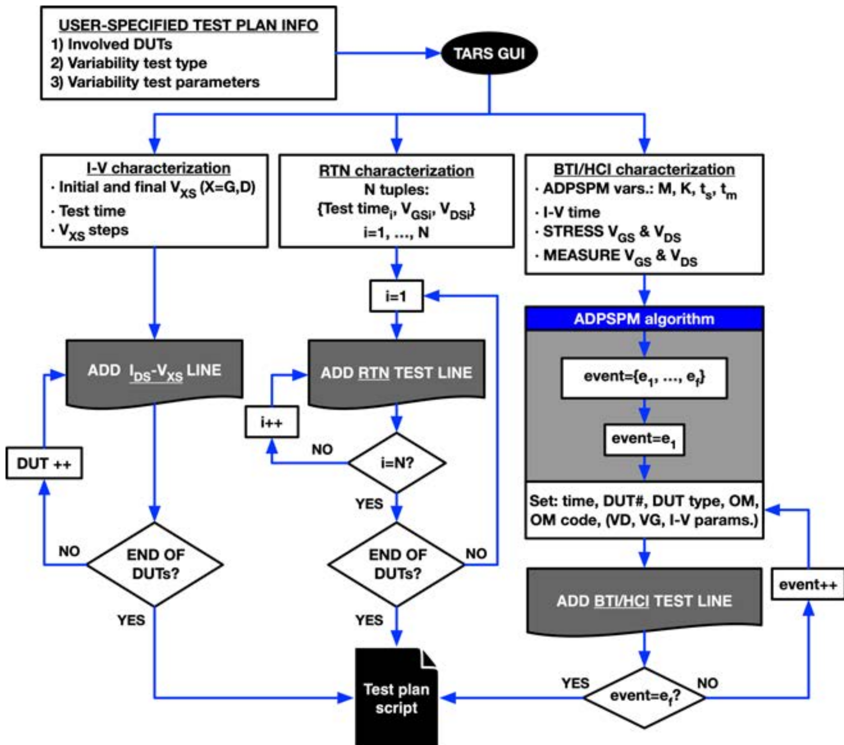


Figure 3.4: Flow diagram of the test plan generation module.

Table 3.5: RTN, BTI and HCI aging test line column variables definition.

	1	2	3	4	5	6	7
	Test time	DUT#	MOSFET type	OM	VC	V_{GS}	V_{DS}
1	100	51	P	ME	20	-0.5	-0.1
2	10	15	N	ST	11	2.5	2.5
3	100	15	N	ME	22	0.6	0.1
4	10	15	N	ST	11	2.5	2.5
5	100	15	N	ME	23	0.6	0.1

For the generation of the test plan for any BTI/HCI aging test, the ADPSPM algorithm is always executed. For instance, Table 3.6 shows an excerpt example of the first 12 HCI test plan lines for three DUTs. In the example, line #1 sets the first device in stress mode for one second. Then, the device must undergo a 100s measurement period. But 1s before the measurement stage finishes, DUT#2 is set in stress (line #3) for 0s. Before the measurement of DUT#1 finishes, the last second of the measurement time is utilized to execute the I_{DS}–V_{GS} curve (line #4) and immediately after DUT#1 is set back to standby (line #5) ending the first cycle of the DUT#1. It is important to notice that to obtain the parallel scheme achieved by the ADPSPM algorithm, line #3 sets DUT#2 in stress mode before the execution of the final $I_{DS} - V_{GS}$ curve measurement of DUT#1 in order to save time. In line #5, DUT #1 is set to standby mode until the next SM cycle starts.

It is important to remark that since no action is taken before the first stress time, the system will wait the exact first stress time as shows in Table 3.6 line #1. On the contrary, after the first stress period, other actions are taken by the algorithm and the subsequent stress time periods are set with 0s because the DUTs have to only be digitally set to stress while the progression of the parallel algorithm will cover exactly the stress time of each DUT. The same procedure is sequentially applied to the rest of devices. The application of the ADPSPM algorithm has achieved two goals: maximum efficiency as the instruments are taking measurements during the maximum possible time, and short and equal delay between stress and measurement.

Figure 3.5 shows the time evolution of the complete aging test sequence defined in Table 3.6. As shown in Figure 3.5, for DUT #1 an initial stress phase is applied followed by a measurement period, i.e., I_{DS}–t and I_{DS}–V_{GS}. Before the completion of each measurement phase the next DUT it set to stress before the end of the previous DUT measurement concatenating the measurement periods of all DUTs saving time.

Table 3.6: ADPSPM HCI aging test excerpt example.

	1	2	3	4	5	6	7	8	9
Test line	Test time	DUT#	MOSFET type	OM	VC	Biasing conditions			
1	1	1	N	ST	11	2.5	2.5		
2	99	1	N	ME	23	0.4	0.1		
3	0	2	N	ST	11	2.5	2.5		
4	1	1	N	ME	33	0.1	0.01	1.2	0.01
5	0	1	N	SB	00	0	0		
6	99	2	N	ME	23	0.4	0.1		
7	0	3	N	ST	11	2.5	2.5		
8	1	2	N	ME	33	0.1	0.01	1.2	0.01
9	0	2	N	SB	00	0	0		
10	99	3	N	ME	23	0.4	0.1		
11	1	3	N	ME	33	0.1	0.01	1.2	0.01
12	0	3	N	SB	00	0	0		

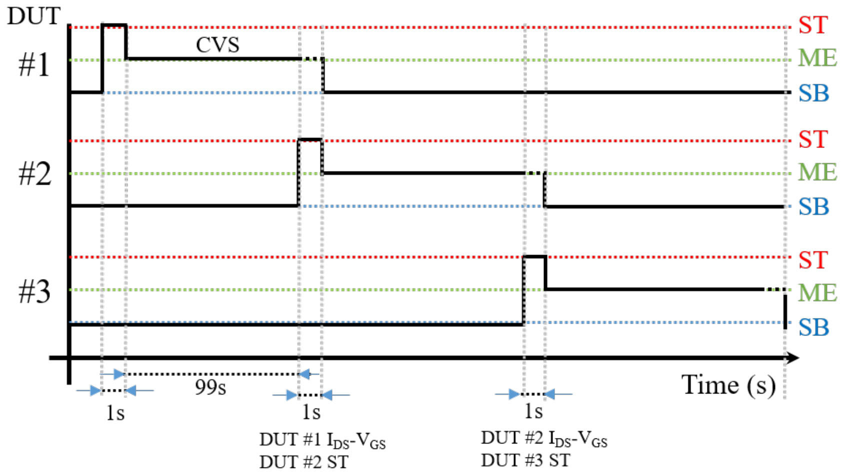


Figure 3.5: Time evolution of the example test proposed in Table 3.6

3.2.2 Elementary test instructions generation module

Each line of the previously described test plan script may involve different setting conditions that have to be sequentially established, e.g., in an I_{DS}–V_{GS} curve, the V_{GS} is swept from 0V to 1.2V and each sweep value will consist

in a single test plan line. Therefore, this second module transforms the test plan script into a set of sequential elementary instructions. Table 3.7 shows the elementary instruction script obtained from the I_{DS}–V_{GS} test plan line detailed in Table 3.3 and the first 4 test lines of the ADPSPM detailed in Table 3.6. Each elementary instruction contains the information to set instruments and DUTs at the correct configuration.

As described in Table 3.7, the elementary test file consists in a fixed 7-column file where $I_{DS} - V_{GS}$ and $I_{DS} - V_{DS}$ I-V test plan lines are split into elementary test lines. In this example, the V_{GS} voltage is swept from 10mV to 1.2V with a step of 10mV while the drain voltage is fixed to 100mV. In the case of an ADPSPM aging test example in Table 3.7, it is important to mention that for BTI and HCI stress or measurement phases, the elementary test instructions are defined with a single test line while the I_{DS}–V_{GS} curve characterization is defined with as many elementary instructions as V_{GS} sampling points are required. All these variables will later be used to construct the GPIB commands that will be sent to the instrumentation equipment together with all required digital IC control commands.

Table 3.7: $I_{DS} - V_{GS}$ test line column variables definition.

1	2	3	4	5	6	7
Elapsed time	DUT#	MOSFET type	OM	VC	V_{GS}	V_{DS}
$I_{DS} - V_{GS}$ Elementary test instructions.						
0.000e-2	15	N/P	ME	33	0.01	0.1
1.000e-2	15	N/P	ME	33	0.02	0.1
2.000e-2	15	N/P	ME	33	0.03	0.1
3.000e-2	15	N/P	ME	33	0.04	0.1
...						
Aging ADPSPM Elementary test instructions.						
0.0000e+0	1	N/P	ST	11	2.5	2.5
1.0000e+0	1	N/P	ME	23	0.5	0.1
1.0000e+2	2	N/P	ST	11	2.5	2.5
1.0000e+2	1	N/P	ME	33	0.01	0.1
1.0100e+2	1	N/P	ME	33	0.02	0.1
1.0200e+2	2	N/P	ME	33	0.03	0.1
...						

3.2.3 Test execution protocol and output test files

The test execution protocol module, illustrated in the flow diagram in Figure 3.6, reads and executes the elementary test instruction file line by line, extracting the test variables and generating the necessary GPIB commands and digital chip control signals to execute the generated test plan.

The test execution protocol first analyses if the currently selected DUT must be changed. If this is the case, the module sends the corresponding row and column matrix addresses of the new DUT to the serial chip interfaces. Then, it activates the selected operation mode (OM) by connecting the drain and gate DUT terminals to the required analog IC paths. The same operation that is performed if the DUT is not changed but a new OM code must be applied to the current DUT. After activating the appropriate DUT and setting the OM, the test execution module obtains the V_{GS} and V_{DS} voltages and sends the appropriate commands to start a stress or a measurement phase:

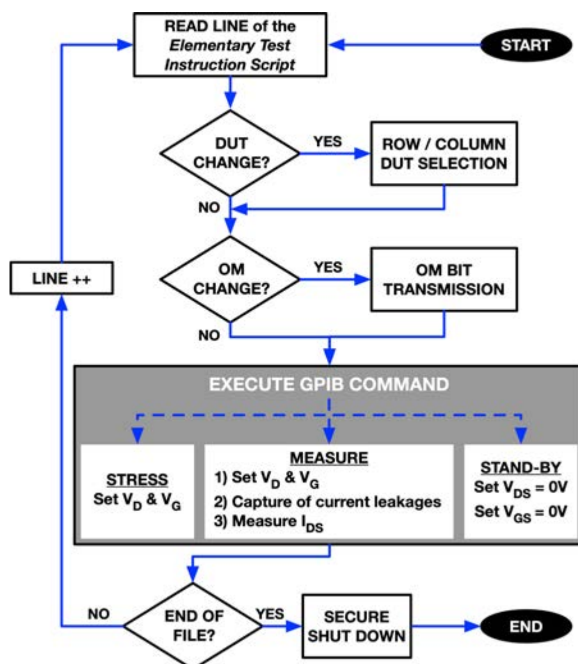


Figure 3.6: Flow diagram of the test execution protocol.

(i) Stress phase (ST): the stress biasing voltages are set to the drain and gate SMUs and will be kept until the end of the test ensuring that all involved DUTs are stressed under the same conditions.

(ii) Measurement phase (ME): the biasing voltages of this mode can be changed each time a DUT is set into the measurement OM because only one DUT will be measured at a time.

For example, when $I_{DS} - V_{GS}$, RTN and BTI/HCI tests are specified for 100 DUTs, a test plan with 2,500 lines is generated. Execution of this test plan produces 20,800 elementary test instructions that trigger the same number of GPIB commands, as well as 2,300 IC digital control instructions. After the set of test experiments specified in the test plan are completed, a secure shut down protocol is executed to prevent any electrical damage to the IC. First, a general RESET signal is sent to all unit cells in order to set the secure standby operation mode for all DUTs; second, all SMUs are disconnected; third, the PCB biasing voltage set by the power supply is removed, and, finally, the GPIB connections with all measurement equipment are terminated.

The test execution protocol performs the measurement phase in 4 steps to ensure minimum time gap between ST and ME phases and to compensate the current leakage of the IC for the particular test conditions:

1. PAD biasing: the algorithm sets the required bias voltages to drain and gate measure lines chip pads (DMF, DMS and GM in Figure 3.12).
2. Leakage characterization: a leakage current measurement is performed through the DMF chip pad to calibrate the DUT current measurement.
3. I_{DS}-t measurement: the current through the measurement path of the chip is captured while before changing the DUT from a previous operation mode, i.e., SB or ST, to the measurement mode.
4. Digital DUT change to measurement mode: after starting the measurement, the OM code of the unit cell DUT is digitally changed to ME mode and the Drain/Gate DUT terminals are connected to the measurement paths. This switching procedure ensures that the time gap between changing to the measurement mode will remain minimum and equal for all DUTs involved in the test.

To illustrate the SM sequence described during an aging test, Figure 3.7 shows the current behaviour of an ENDURANCE IC DUT when switching from stress to measurement modes. The first measurement points (100ms) represent only leakage because the DUT is still connected to the stress lines,

after the leakage current measurement is done, the DUT OM is set to the measurement mode and the registered current changes from the pA to the μA with a time step of 2ms. The figure also shows how during the measurement time window of 100s the transistors experience a V_{th} recovery as can be observed by the abrupt current jumps present in the current trace.

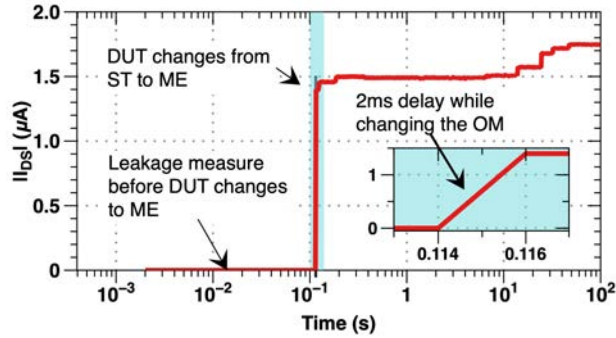


Figure 3.7: $I_{DS} - t$ trace measurement after BTI stress executed following the 4-step process when switching from ST to ME with constant biasing of $V_{GS} = 500mV$ and $V_{GD} = 100mV$.

During the execution of a variability test two files are created: The output measurement file and a measurement log file. The output measurements file, where an excerpt of this file is shown in Figure 3.8, consists in a plain text file organized in 9 data columns with all the data extracted by the measurements system during the execution of the test for further analysis. All measurements executed to a DUT are stored in a single output measurement file containing the following data from column #1 to #9: DUT #, I_{DS} measured current, the measurement time, the applied V_G, the applied V_D, the corresponding elementary test file line for the executed measurement, the MOSFET type, the OM and finally, the VC.

In parallel to the on-line creation of the output measurement file, the TARS software main engine creates a Measurement log text file where every action executed during a test is saved with the system date and time that can be seen in real time in the main GUI of the software. This allows to precisely analyse the integrity of all the executed tests in terms of phases execution and elapsed times between the different phases of an ADPSPM aging test parallel execution. Figure 3.9 shows the initial lines written in a Measurement log file for each test.

```
1  DEV Current      time     VG   VD   TIMMING_LINE     MOS_type     OP_MODE NUM_OP_MODE
2  298 3.000000e-10     6.2828984336e+02     1.500000e-02     1.000000e-01     8586     N    ME   33
3  298 2.550000e-10     6.2837795008e+02     2.697000e-02     1.000000e-01     8587     N    ME   33
4  298 2.900000e-10     6.2846580657e+02     3.893900e-02     1.000000e-01     8588     N    ME   33
5  298 2.800000e-10     6.2855376743e+02     5.090900e-02     1.000000e-01     8589     N    ME   33
6  298 1.900000e-10     6.2864195156e+02     6.287900e-02     1.000000e-01     8590     N    ME   33
7  298 2.000000e-10     6.2872999069e+02     7.484800e-02     1.000000e-01     8591     N    ME   33
8  298 1.450000e-10     6.2881793016e+02     8.681800e-02     1.000000e-01     8592     N    ME   33
9  298 9.500000e-11     6.2890587605e+02     9.878800e-02     1.000000e-01     8593     N    ME   33
10 298 1.500000e-11     6.2899387070e+02     1.107580e-01     1.000000e-01     8594     N    ME   33
11 298 2.150000e-10     6.2908188930e+02     1.227270e-01     1.000000e-01     8595     N    ME   33
12 298 2.850000e-10     6.2916992415e+02     1.346970e-01     1.000000e-01     8596     N    ME   33
13 298 6.600000e-10     6.2925781657e+02     1.466670e-01     1.000000e-01     8597     N    ME   33
14 298 1.020000e-09     6.2934589377e+02     1.586360e-01     1.000000e-01     8598     N    ME   33
```

Figure 3.8: Example of a measurement output file.

Each line gives information about the current action executed by the software together with the system date and time. Once the initial configuration of the test, e.g., measurement folder creation (line #1), instrument connections (digital instructions module connection at line #3), the test execution is reflected in the Measurement log file by a set of 3 log lines which are repeated each time that an action is taken by the TARS software (see lines #10 to #12):

```
1  New output data folder created!; 07-Jun-2018 13:47:18
2  "Timing experiment flow" text file correctly created: 07-Jun-2018 13:47:28
3  Connection stablished wiht the DAQ-6501: 07-Jun-2018 13:48:15
4  Power supply connected and PCB biased: 07-Jun-2018 13:48:16
5  Keysight B1500A GPIB connection established: 07-Jun-2018 13:48:16
6  NMOS AND PMOS RESETS DISABLED: 07-Jun-2018 13:48:56
7  Starting the PMOS matrix STANDBY operation mode setting: 07-Jun-2018 13:48:58
8  Ending the PMOS matrix STANDBY operation mode setting: 07-Jun-2018 13:53:51
9  The test will end at: 08-Jun-2018 17:05:52
10 Testing DUT # 301; OP MODE --> ME; Code: 33; VG: 0.015; VD: 0.1; 07-Jun-2018 13:53:52
11 EQUIPMENT AND INSTRUCTION SYNCHRONIZATION (DELAY): 0s 07-Jun-2018 13:53:52
12 ACHIEVED PERCENTAGE OF THE MEASURE: 0% 07-Jun-2018 13:53:52
13 Testing DUT # 301; OP MODE --> SB; Code: 0; VG: 0; VD: 0; 07-Jun-2018 13:53:59
14 EQUIPMENT AND INSTRUCTION SYNCHRONIZATION (DELAY): 2.2911s 07-Jun-2018 13:53:59
15 ACHIEVED PERCENTAGE OF THE MEASURE: 0.01% 07-Jun-2018 13:53:59
```

Figure 3.9: Example of a Measurement log file.

- The tested DUT line information includes: the DUT number, the OM mode defined and the V_G/V_D voltages applied to the DUT (see line #10).
- The laboratory synchronization delay: total accumulated delay time due to code execution and instruments communication (see line #11).
- Real-time achieved percentage of the executed test (see line #12).

3.3 Hardware characterization framework

The first step for a trustworthy transistor characterization is based on a reliable hardware framework for accurate and individual device test. In this regard, Figure 3.10 shows a sketch of the equipment used in the characterization framework implemented for ENDURANCE DUT IC characterization with the following list of items:

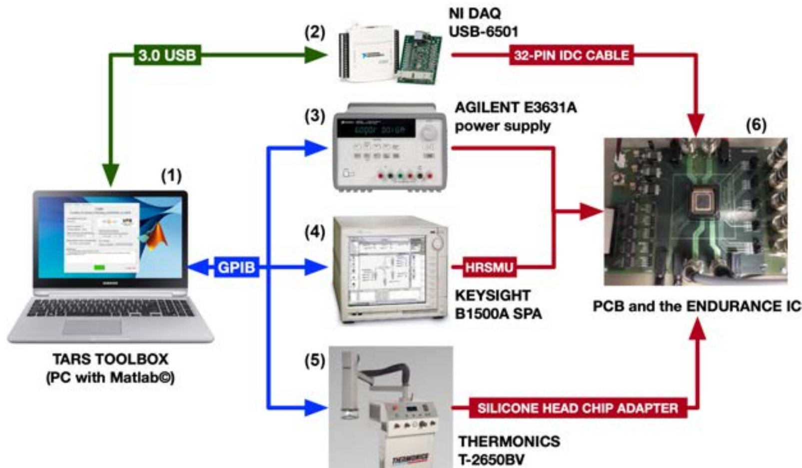


Figure 3.10: Schematic representation of the laboratory hardware setup for ENDURANCE IC characterization.

1. A personal computer (PC) equipped with Matlab® software where the TARS software is executed.
2. A USB Digital Acquisition System (DAQ), model USB-6501 from National Instruments, equipped with 24 digital I/O channels [140].
3. The Agilent E3631A power supply for PCB (and subsequently IC) biasing using a 5V/1A DC output voltage [141].
4. The Keysight Semiconductor Parameter Analyzer (SPA) model B1500A. This equipment has been provided with 4 Sense Measurement Units (SMU), with Force-&-Sense outputs intended for precise voltage biasing and low current measurement [142]. Voltage and current precise application can be applied to the tested devices in constant or sweep traces during device testing.

5. The T-2650BV Thermonics precision temperature system with a temperature ranging from room temperature to 120°C [143] that utilizes an air flux for heating or cooling the chip by means of a thermopar to control the temperature of the chip.
6. A full-custom printed circuit board (PCB), where the ENDURANCE IC is inserted using a zero insertion force socket[4] model JLCC68.

As depicted in Figure 3.10, the digital communications between PC and all laboratory measurement equipment is performed by means of GPIB bus and the USB DAQ model 6501 from National Instruments in order to stablish a precise communications between all laboratory instruments. Moreover, Figure 3.11 shows the final characterization setup assembled in the laboratory to test the array-based ENDURANCE IC chip. A critical concern consists in protecting the ENDURANCE IC and the PCB against electrical damage coming from human-body EDS [144]. This kind of protection is a priority since the stress PADs of the IC (i.e., DSF and GS) are not EDS protected in order to reduce its series resistance and, consequently, the associated voltage drop. In this regard, an anti-static mat and wrist strap have been used in the laboratory to isolate all hardware setup components from the electric damage caused by ESD.

Figure 3.11: Laboratory characterization setup for variability testing corresponding with the schematic drawing shown in Figure 3.10.

[4] A type of IC socket that requires little force to insert the IC chip. Moreover, all the IC PADs are contacted at the same time and with the same pressure.

3.3.1 Printed circuit board for IC characterization

One of the most important elements of the characterization setup is the PCB that permits the insertion and extraction of different ENDURANCE IC samples. The PCB also harbours all the necessary connections, like the triaxial connectors for testing and current measurements or digital IC control. An accurate design of the PCB provides access to the IC chip PADs and preserve the accuracy of the measurements.

Figure 3.12 shows the top view of the PCB for the ENDURANCE IC chip testing. It has been divided in three sections: the input digital block (Figure 3.12, section A), the biasing circuitry block (Figure 3.12, section B) and the socket and connectors block (Figure 3.12, section C). These blocks are explained in more detail below:

A Input digital block: this block (Figure 3.12, section A) is in charge of receiving the digital signals for IC control sent from the DAQ 6501. All digital lines are connected to the PCB through a 32-pin Insulation-Displacement connector [5] (IDC). A set of 5 electromagnetic isolators model ADUM1400 [145] are used between the DAQ outputs and the digital IC chip inputs to isolate the IC chip PADs from external noise and to shift all digital signals from the TTL 5V/0V output DAQ level to the 2.5V/0V input voltage level of the digital IC chip pads.

B Biasing circuitry block: Three precise and stable biasing voltages are required for the ENDURANCE IC chip biasing: the biasing of the digital core circuits (1.2V), the digital input/output (I/O) pads (2.5V) and the transmission gates that enable stress voltages up to 3.3V. For this purpose, a dedicated biasing circuitry that uses three CMOS high precision voltage regulators have been implemented in the PCB circuitry (Figure 3.12, section B). This circuitry provides, from a single 5V-1A DC power supply voltage input, the three stable IC biasing required for IC circuit operation.

C Socket and connectors block: This part of the PCB (Figure 3.12, section C), includes a 68-pin zero-insertion force socket model JLCC68 that allows easy replacement between different ENDURANCE chip samples.

Analog pads of the chip are connected to 9 triaxial connectors to supply voltages from the Keysight B1500 SPA and simultaneously, measure

[5] An IDC connector, also known as insulation-displacement contact, is designed to work with ribbon cables, and uses sharp contacts to clamp onto the wire.

current. One of the most critical signal paths of the IC chip corresponds to the four Drain Measure Force PADs of the IC DUT submatrices (i.e., DMF_NL, DMF_NR, DMF_PL and DMF_PR). Thus, in the PCB design, the 4 DMF connections have individual triaxial PCB connectors to connect directly the SPA to one of the IC submatrices. Moreover, a built-in guard has been included in the PCB design starting from each DMF connection and ending at the corresponding IC chip socket pin to prolong the guard of the triaxial cables to minimize voltage drop and noise when biasing and measuring current during variability tests.

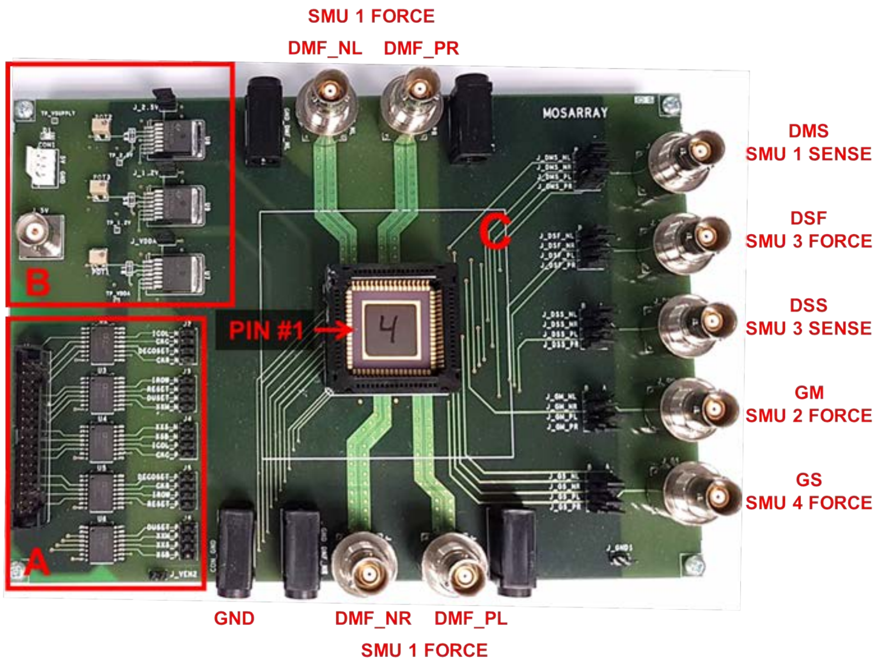


Figure 3.12: Top view of the fabricated PCB for ENDURANCE IC test.

The remaining triaxial connections located at the right side of the PCB (Figure 3.12, section C) named DMS, DSF, DSS, GM and GS do not have a built-in guard either because the current is negligible (which happens with the gate and sense lines) or because they are used for stress purposes where the current is not measured in most of the test scenarios. These signals are physically multiplexed along the four DUT matrices, so that only five additional triaxial connectors are necessary.

Gate and Drain standby pads are physically connected to the analog ground or analog power supply voltage, i.e., N- or P-type transistors, respectively. Finally, as it can be seen in the center of the PCB, a squared area surrounding the IC socket has been left free of connectors and electronic components because, it will be used to place the Thermonics silicone head adapter over the IC chip, so that the temperature of the chip can be established and monitored.

The design of the described PCB implies that only one of the four IC DUT matrices can be physically selected to conduct tests. As shown in Figure 3.12, section C, a single triaxial connector has been provided for the DMS, DSF, DSS GM and GS signal so they have to be multiplexed and, for this, a set of 20 jumpers have been included in the PCB design to manually multiplex those analog signals to select the DUT of the desired matrix of the chip.

Figure 3.13 shows the 20-jumper placement for the individual selection of each matrix inside the ENDURANCE chip. To physically select one of the four DUT matrices by means of the provided PCB jumpers, three pins define if a chip PAD is short-circuited to the triaxial connector signal or if it is biased with standby voltage, i.e., VSS for NMOS submatrices and VDDA for the PMOS ones. To conduct a signal from the triaxial connector to the IC PAD, a jumper must be placed between the central and the labelled 'A' pins. On the contrary, if the chip PAD had to be biased with standby voltage, a jumper should be connected between the central and the labelled 'B' pins. Moreover, in order to avoid unwanted voltage drops at the non-selected DUT matrices, a banana plug is used to bias with the required standby voltage the DMF PADs during the measurements.

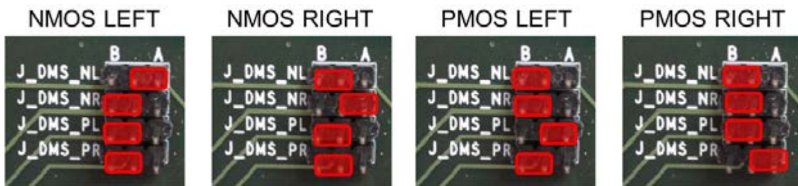


Figure 3.13: Example of the PCB jumper collocation for analog signals multiplexing.

3.3.2 IC leakage current measurement

When studying MOSFET variability phenomena, two different types of accelerated stress can be applied to the samples to reduce the total testing time: the use of overvoltage biasing conditions and/or the application of high temperature to the DUTs. An important consideration when dealing with IC chips for device variability testing, is that the application of high temperature not only affects the phenomena linked to the DUT itself, the entire IC architecture can be also affected by temperature. In this regard, any device or circuit placed between the DUT and the external measurement source can alter the measured current and finally, convert the measurements into unreliable data.

In order to avoid any measurement artefact due to temperature gradient when measuring DUTs of the ENDURANCE IC, two current measurements have been executed by means of a physical chip and the complete measurements setup, to characterize current behaviour under fixed voltage measurements and temperature stress application. Figure 3.14 shows two IC current characterizations conducted under the same temperature stress pattern: first, the drain current obtained from an nMOS transistor with dimensions: $W = 1\mu m$ and $L = 1\mu m$ (Figure 3.14 (a)) and second, the current captured without connecting any DUT into the measurements path of the IC chip. This current corresponds to the contribution of all sources of leakage current present in the measurement channel of the ENDURANCE IC (Figure 3.14 (b)).

For both current characterizations, a constant voltage biasing of $V_{GS} = 500mV$ and $V_{DS} = 100mV$ has been set externally to the IC measurement PADs, i.e., DMF, DMS and GM respectively. The voltage was kept constant during both tests by means of the B1500 SPA while constantly capturing the measurement channel current. The Thermonics precision temperature system was used to apply externally to the IC chip a rising and falling temperature waveform in a 5-step test process:

1. The temperature is kept at $25°C$ for 1 minute.
2. The temperature is raised from $25°C$ to $110°C$ by $30°C$/minute.
3. The temperature of $110°C$ is kept for 1 minute.
4. The temperature is lowered from $110°C$ to $25°C$ by a $30°C$/minute.
5. The temperature is kept at $25°C$ for 1 minute.

As shown in Figure 3.14 (a), the DUT drain current measured increases as well as the temperature. Nevertheless, as the applied temperature reaches $\approx 75°C$, the DUT current starts decreasing instead of continuously rising as expected as shown in Figure 3.14 (a) between 200s and 300s. The opposite behaviour can be observed between 350s and 450s when lowering the temperature from $110°C$ to $25°C$. This drain current behaviour is understood by an artifact produced by the leakage current present in the measurement channel as demonstrated by the red trace shown in Figure 3.14 (b). The captured leakage current corresponds to the contribution of all leakage sources present in the IC which are enlarged and reduced during the application of positive and negative temperature ramps.

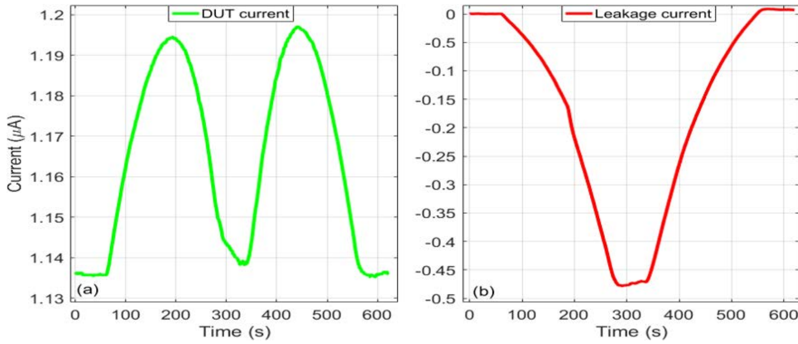


Figure 3.14: Characterization of the current present in the measurement path of the chip under a temperature stress pattern ($25°C$-$110°C$-$25°C$). (a) DUT drain current measurements of an nMOS transistor connected to the IC measurement path and (b) measured current through the measurement path without connecting any DUT to the IC measurement path.

To clarify the behaviour of the IC current during the tests conducted in Figure 3.15 (a) and (b), shows the evolution of the current in both tests as a function of the applied temperature. As shown in Figure 3.15 (a), when the temperature reaches $\sim 70°C$ the leakage current achieves a value which forces the DUT current to stop rising and start decreasing (green trace) because of the negative value of the increasing leakage current (red trace). The sign of the leakage current is negative because the current flow goes from the DMF (cathode of the EDS diode) to the VSS (anode of the ESD diode) in the IC. The opposite DUT current behaviour can be recognized when lowering the temperature from $110°C$ to $25°C$ because the DUT current increases and the leakage current decreases. In order to achieve only the DUT drain current

during the described test, Figure 3.15 (b) shows the DUT current were the leakage current has been cancelled from the trace. Finally, the behaviour of the DUT current rises as well as the IC temperature does and decreases in the opposite situation.

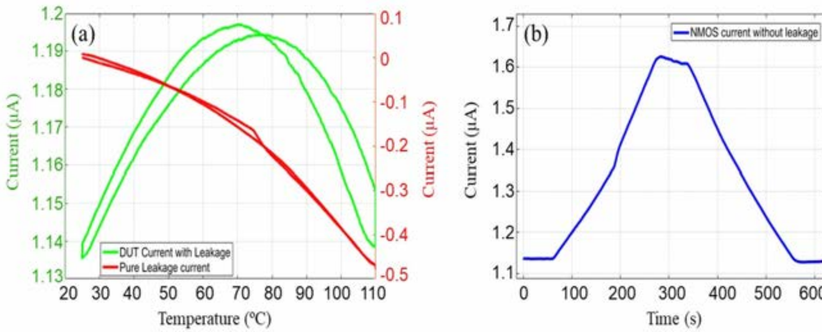


Figure 3.15: In (a) behaviour of the DUT current with leakage together with the pure leakage current under temperature stress and (b) shows the on-transistor current characterization with full measurement channel leakage current cancellation following the temperature sequence of 25°C - 110°C - 25°C.

As has been demonstrated, the leakage current present in the DUT current measurements, by means of an IC for statistical device testing, is not negligible and is also strongly dependent to the temperature conditions. In this regard, any measurements conducted utilizing the ENDURANCE IC chip is accompanied by an initial leakage current measurement before connecting digitally the DUT to the measurement IC paths. Moreover, this leakage current measurement will be further used to compensate all leakage sources present in the measurement channel from the DUT measured current as shown in Figure 3.7 (b).

3.4 TARS software: graphical user interfaces

With the aim of having full control of the described flexible setup for variability phenomena characterization, three individual user-friendly GUIs have been designed to define massive I-V, RTN and BTI/HCI aging tests while maintaining a perfect synchronization between the hardware and the TARS

software. The main GUI, named ”Test setup and monitoring GUI”, the second GUI, named ”Test File Generation GUI”, utilized for the automatic generation of massive test plan files and, finally, the ”data post-processing GUI” intended for visualizing and processing measured data. Each one of the GUIs is designed for a specific functionality purpose:

1) The test setup and monitoring GUI: this GUI, whose graphical design is shown in Figure 3.16, corresponds to the main control GUI of the TARS software that allows users to select and execute variability experiments by means of a set of text boxes and drop down menus. The following options are accessible to the users:

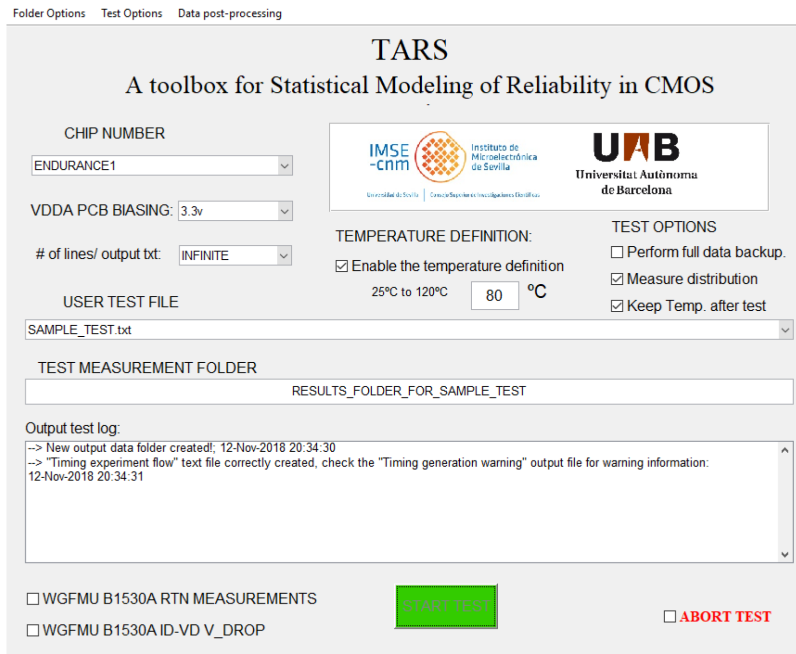


Figure 3.16: The TARS main test setup and monitoring GUI.

- Chip number selection: selection of the ENDURANCE IC chip number inserted in the PCB socket.
- Analog voltage PCB biasing: selection of the VDDA voltage biasing: 2.5V or 3.3V.

- Temperature definition: PCB temperature control during test execution with a range between $25°C$ and $120°C$.
- User Test File: with this drop down menu, users can select any previously created test plan stored in the selected chip master folder.
- Test measurement folder: definition of the test name. Together with the entered name, the software concatenates the system date to make each test name unrepeatable.
- Test options: test options enabled or disabled by the user before the test execution, like the execution of the finished measurement backup or the option to perform a full backup of all the stored measurements of all ICs in the master folder.
- Output test log window: live window where every action executed by the software during variability test is shown. As shown in Figure 3.16, the first action executed by the TARS software corresponds to the creation of the measurement folder and the automatic generation of the elementary test instruction file before starting the selected "SAMPLE_TEST.file". Each line shown in the output test log window is stored in the Measurement log file (see Figure 3.9).

In order to endow more versatility to the Test setup and monitoring GUI, an upper ribbon with three drop down menus have been included in the design to set additional test options as shown in Figure 3.17:

- Folder options: this menu allows to open the selected ENDURANCE IC chip master folder where all the executed measurements for each particular IC chip number are stored. It also allows to open the user test plan files folder where all created test plans are stored.
- Test options: this drop down menu allows to launch the "Test plan generation" GUI, change the sampling rate of the measurement phases for BTI/HCI aging and RTN phenomena tests, compute the extraction of the V_{th0} and μ_{eff} as will be explained in chapter 4 or execute the creation of a new ENDURANCE test folder when a new chip is inserted in the PCB for testing.
- DUT post processing: this GUI permits users to easily and quickly plot any stored variability test. As shown in Figure 3.17, users can select plotting option like I-V measurements, RTN or to process the results of the concluded aging tests.

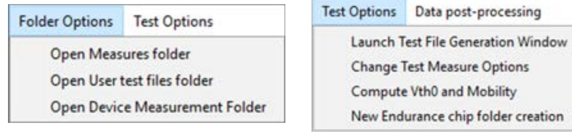


Figure 3.17: Additional main GUI menus.

2) The test plan generation GUI shown in Figure 3.18: using a set of user-friendly an intuitive text and check boxes, the users are able to select DUTs or macrocells for testing, defining the variables of the available test to be executed and generating, in a few milliseconds, a complete test plan file. The available test options are: I-V measurements, i.e., $I_{DS} - V_{GS}$ or $I_{DS} - V_{DS}$, RTN characterization, or the generation of BTI/HCI aging test plan files:

- I-V measurements: controlled by a check box, this section of the GUI allows to include in the test plan the $I_{DS} - V_{GS}$ and/or $I_{DS} - V_{DS}$ curve characterization. As shown in Figure 3.18, users can activate one of both curve characterizations and all parameters of the curves are user-defined, e.g., the curve testing time, V_{GS} and V_{DS} ranging voltages or the number of V_{GS} steps when the $I_{DS} - V_{DS}$ curve characterization is selected.

- RTN characterization: in this case only 3 variables should be introduced in the text boxes if the RTN generation check box is selected: the RTN test time in seconds, the V_{GS} and the V_{DS} voltages. If more than one RTN test must be executed with different test variables, those variables have to be entered separated by commas as shown in the RTN section in Figure 3.18, where 3 RTN tests have been defined.

- BTI/HCI aging test plan definition: allows the definition of the necessary parameters to perform BTI/HCI tests like the number of SM cycles, the K parameter (the exponential base for the increasing stress time periods), the initial stress time and the fixed measurement phase time and the required voltages applied during the stress and the measurement phases of the aging test. The algorithm of the test plan generation utilizes the novel ADPSPM technique developed in this thesis.

3) The Data Post-Processing GUI: allows an easily visualization of the output results from any completed test. The GUI permits to select between the different chip folders and measurement folders stored in the particular

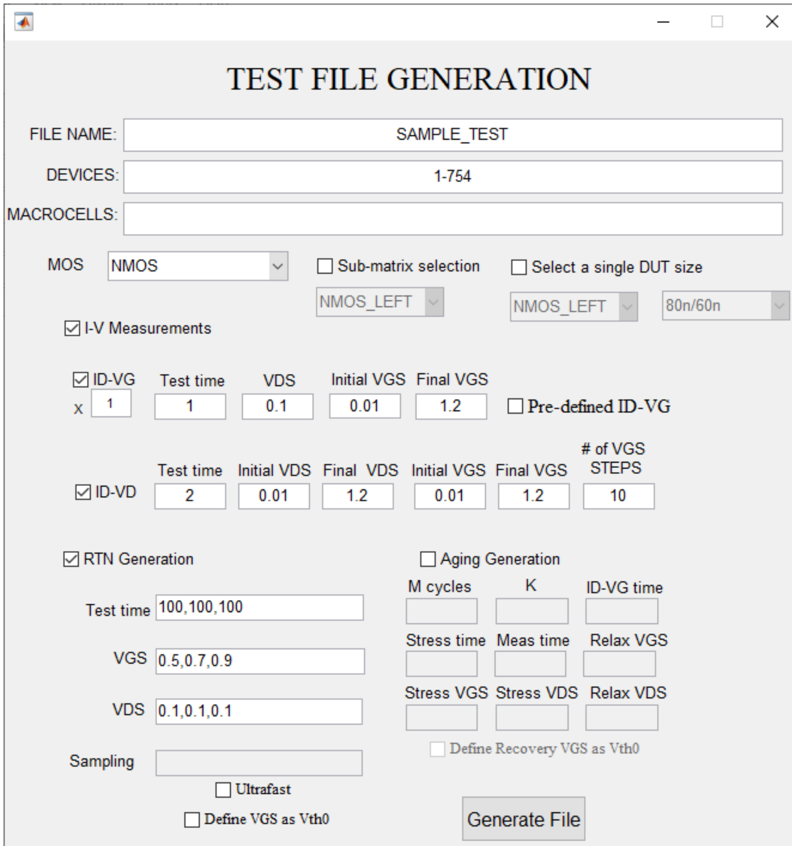


Figure 3.18: The Test File Generation GUI of the TARS software.

chip folder selected. For instance, Figure 3.19 shows a set of 784 $I_{DS} - V_{GS}$ curves extracted from an ENDURANCE IC chip NMOS right matrix. The plotted $I_{DS} - V_{GS}$ curves corresponds to the 8 different transistor geometries measured in the IC chip matrix.

To visualize the results of any selected test, one of the three options can be activated: plotting the stored I-V measurements, plotting the RTN traces obtained from an RTN test experiment and finally, starting the CMOS Parameter Extraction option. The latter option initiates an algorithm that obtains the degradation of the transistor's parameters after the BTI and/or HCI aging tests that can be later used to feed the stochastic aging models like the PDO model [98].

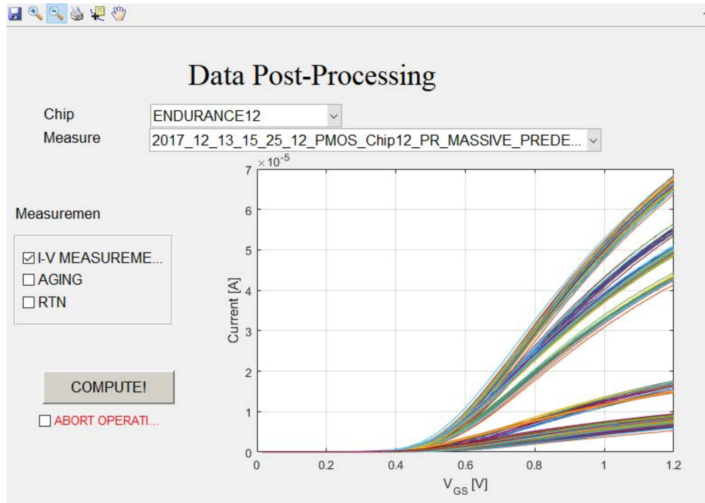


Figure 3.19: The Data Post-Processing GUI of the TARS software.

3.5 Conclusions

In the present chapter, a complete measurement framework for the characterization of CMOS transistors with array-based ICs has been described. Two main parts compose the described measurements setup: the software part consisting in a full-custom user-friendly software, named TARS, and a complete hardware laboratory instruments focused on achieving a trustworthy MOSFET device variability characterization level.

The system has been used to characterize the 65-nm transistors of the ENDURANCE IC, a novel IC design that permits the execution of TZV and parallel stress TDV characterization, including HCI aging phenomena. The measurement setup enables several tests: ramped voltages ($I_{DS} - V_{GS}$, $I_{DS} - V_{DS}$), RTN measurements, and BTI and HCI aging evaluations. These tests can be defined in a matter of seconds while thousands of commands are automatically generated for instrumentation and chip control.

From the hardware design point of view, the main hardware block corresponds to the full-custom PCB utilized to insert the ENDURANCE IC chip for the MOSFET device variability characterization. The PCB has been carefully designed with different built-in guards to prolong the SPA triaxial

cable guards to the IC chip pins in order to minimize leakage fluctuations that can occur in the connection cables utilized for chip biasing and device current measurements. Furthermore, a set of electromagnetic isolators have been included in the PCB design, first, to isolate the grounds of the IC chip from the PC ground to keep all digital signals noise-free and second, to perform an accurate digital voltage shift of the 5V TTL DAQ output voltages to the 2.5V CMOS IC chip levels.

One of the main software goals presented in this chapter is the creation of the novel ADPSPM technique. This aging characterization technique consists in the automatic generation of BTI and HCI aging tests involving hundreds of transistors where a significant test time reduction has been successfully achieved thanks to the parallelization of the device stress phases maintaining serial measurement phases. The ADPSPM algorithm has demonstrated a significant aging test time improvement against other parallel testing technique like the PSPM and the traditional serial device wafer testing.

The main engine of the TARS software automatically controls the execution of the user-defined tests, generating intermediate full-custom files to execute each required test action maintaining a perfect synchronization between all laboratory instrumentations and the IC chip. To simplify the tedious task of manually creating thousands of test lines when statistical device TZV and TDV tests have to be executed, a set of 3 user-friendly GUIs have been designed and described in the present chapter. The GUIs allow to automatically create in a few seconds large test files by the users with all necessary variables to execute any required variability test.

The described measurement setup in combination with the ENDURANCE IC, empowered by the ADPSPM algorithm, allows the application of smart stress parallelization techniques to an unlimited number of devices. The ADPSPM testing technique smartly overlaps the stress phases of the aging tests, reducing the measurement time from years to hours.

CHAPTER 4

Expected MOSFET lifetime prediction and parameters extraction methodologies

In previous chapters, the need of new frameworks intended to execute trustworthy statistical variability phenomena characterization in nanometric transistors has been described for accurate aging degradation phenomena test. The ENDURANCE IC chip and the full-custom TARS software have been presented, both utilized to execute statistical MOSFET variability characterization in very short times. In this scenario, the present chapter will describe the achieved results based on statistical characterization of the ENDURANCE chip samples where, in the first section, the reliability of the tested samples will be studied in terms of expected lifetime prediction under BTI/HCI TDV aging phenomena.

In the second section, three of the novel algorithms developed in the context of this thesis during three different research internships conducted at IMSE-CNM in Sevilla (Spain), at the Karlsruhe Institute Für Technology (KIT) (Germany) and at IMEC research center in Leuven (Belgium) will be presented. The algorithms are focused on the extraction of the transistor's model card parameters that can be used during device/circuit reliability simulations together with a novel implementation approach of the PDO model utilizing GPU-based computational resources. In the fourth and the fifth sections, experimental results obtained from the characterization of an advanced 28-nm array-based IC design and the description of a novel ultrafast characterization algorithm, that can be used for testing devices in the micro/nano second time scale, will be, respectively described.

4.1 Lifetime prediction based on time-zero and time-dependent statistical variability measurements

When dealing with time-zero (TZV) and time-dependent (TDV) variability phenomena, a significant number of devices must be characterized to obtain trustworthy lifetime technology predictions due to the stochastic nature of variability phenomena. In this sense, the present section will show, in the first place, the statistical TZV obtained from the characterization of the ENDURANCE IC chip samples. Later on in this section, transient RTN phenomenon and BTI/HCI aging TDV phenomena will be statistically characterized in order to account for the expected device lifetime prediction.

4.1.1 Time-zero variability measurements

As described in Chapter 1, Time-zero variability, consists of a constant, either random or systematic, shift of intrinsic device parameters, such as threshold voltage (V_{th}) or mobility (μ). In this sense, Figure 4.1 (a) and (b) shows the I_{DS}–V_{GS} characteristics of a total of 784 pMOS devices of a single ENDURANCE IC, where the different current levels achieved by each of the 8 transistor geometries of the ENDURANCE chip are depicted.

As shown in Figure 4.1 (a), a large TZV can be observed from device-to-device measurements, especially for the 80nm/60nm geometry (yellow traces in Figure 4.1 (a) and (b)). Moreover, different current levels are observed, as expected, for each tested geometry as depicted in Figure 4.1 (b).

MOSFET V_{th} and μ parameters are two of the most utilized parameters utilized as degradation metrics to describe device shifts behaviour [146]. Traditional MOSFET models, like the BSIM4 [147] rely on the accuracy of the experimentally extracted transistor's parameters. In this sense, the commonly used parameters extraction methods are utilized with experimental measurements of static MOSFET transfer characteristics utilizes 1st and 2nd order extraction methodologies like, for instance, the constant current (CC) method.

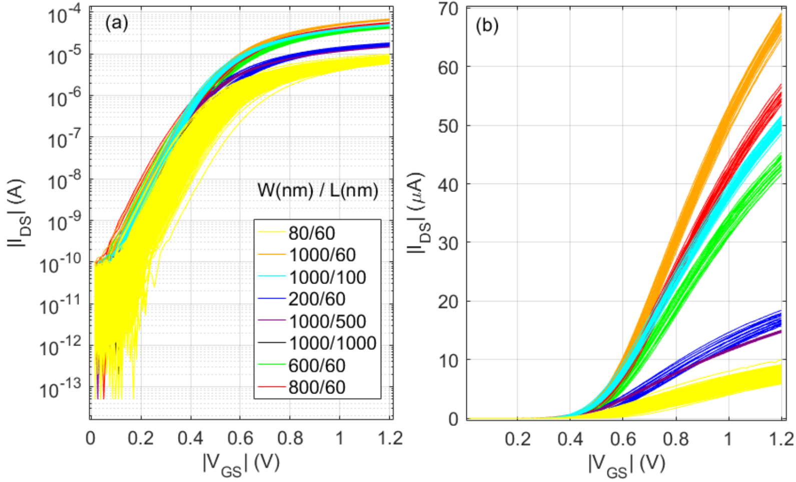


Figure 4.1: 784 PMOS $|I_{DS}| - |V_{GS}|$ curves measured for the eight available geometries in the ENDURANCE IC.

The CC method evaluates V_{th} as the V_{GS} value that corresponds to an arbitrary constant I_{DS} level, with a low V_{DS}, e.g., $\leq 100mV$. The constant I_{DS} value should be scaled to the device width and length dimensions (W/L) for a better accuracy. Nevertheless, the CC method is totally dependent on the chosen I_{DS} value but, on the contrary, it's widely used because of its simplicity [148–150].

Prior to any device TDV test execution, a complete TZV analysis has been executed to the devices of 15 ENDURANCE IC chip samples. In this sense, I_{DS}–V_{GS} curve characteristics have been executed to the nMOS and pMOS on-chip devices to obtain the non-degraded V_{th} distributions by means of the CC extraction methodology.

Figure 4.2 shows the empirical CDFs of the non-degraded V_{th} extracted values of both nMOS in Figure 4.2 (a) and pMOS devices in Figure 4.2 (b) from the 80nm/60nm transistors geometry. From the CDFs shown in Figure 4.2 (a) and (b), no correlation can be observed between the different ENDURANCE chip samples and all chips shows a gaussian distribution of the V_{th} parameter as expected. For the nMOS case, a maximum V_{th} variation between different chip samples is $\approx 16mV$ of in terms of V_{th} (see inset of Figure 4.2 (a)), while for the pMOS case the $|V_{th}|$ total variation across the 15 chip samples correspond to $\approx 23mV$ (see inset of Figure 4.2 (b)).

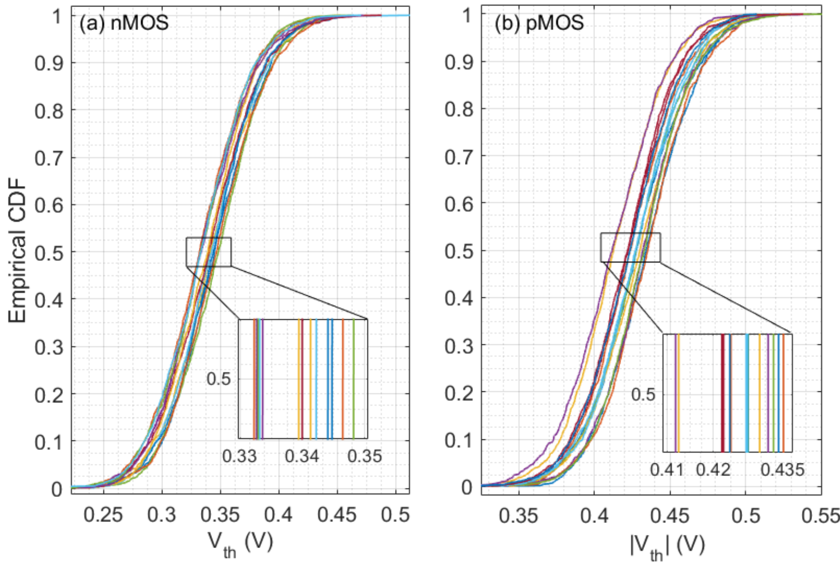


Figure 4.2: Empirical CDFs of the V_{th} parameter for the 80nm/60nm nMOS in (a) and pMOS in (b) transistors of 15 ENDURANCE chip samples.

The CDFs shown in Figure 4.2 (a) and (b) reflects that all tested transistors are operational and the TZV from chip-to-chip has been accurately characterized. Variability of intrinsic transistor parameters plays a special role when dealing with diverse transistor areas. In this sense, Figure 4.3 depicts the behaviour of the transistor's $< V_{th} >$ as a function of the transistors area. The Figure 4.3 displays, for each of the 15 IC chip samples, the $< V_{th} >$ for nMOS transistors, in red dashed lines, and pMOS transistors, in blue dashed lines, together with a thick straight line that defines the $< V_{th} >$ of all data for both transistor types. In Figure 4.3, two distinguished $< V_{th} >$ tendencies for both nMOS and pMOS transistor types can be easily located. In the first case, the $< V_{th} >$ increases when the transistor's length is fixed to 60-nm and the width is increased from 80-nm to 1-μm. In the second case, the transistor's width is fixed to 1-μm while the length is varied from 60-nm to 1-μm, in this case the $< V_{th} >$ decreases as shown in the figure.

As postulated by Pelgrom's rule [2], MOSFET $\sigma(V_{th})$ variability is inversely proportional to the square root of the transistor area. However, in nowadays ultra-scaled area devices $\sigma(V_{th})$ mismatch does not always follow this straightforward rule [151]. In this sense Figure 4.4, shows the typical Pelgrom plot where the $\sigma(V_{th})$ is plotted against $1/\sqrt{W \cdot L}$, where more than

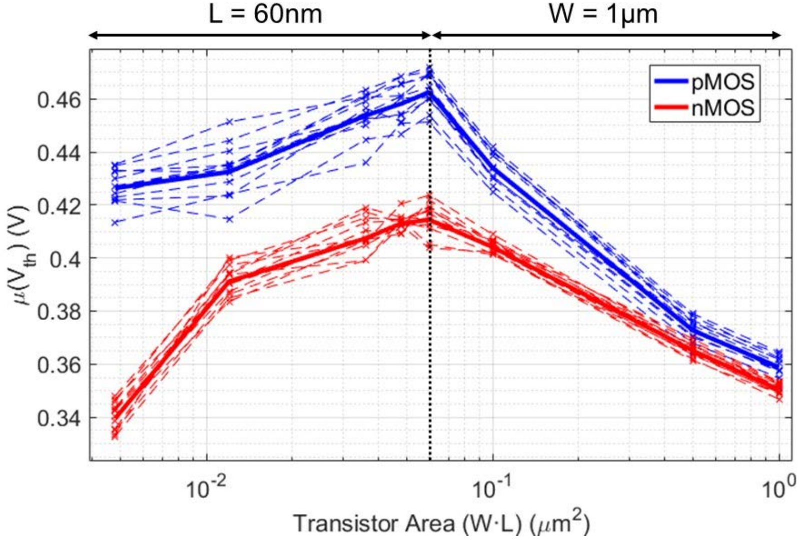


Figure 4.3: $< V_{th} >$ plot as a function of the transistor area for nMOS and pMOS transistors from 15 different ENDURANCE chip samples.

40.000 devices have been tested. A deviation from Pelgrom's rule is depicted in Figure 4.4 by breaking the $\sigma(V_{th})$ variability dependence in two different regions as shown in the figure by a black dashed line: to the left, the length of the devices is kept constant L_{CONST}, while at the right side, the width of the tested devices is kept constant W_{CONST} while the length is varied.

As shown in the Figure 4.4, individual linear fittings have been executed for nMOS and pMOS statistical data, separately for L_{CONST} and W_{CONST} regions showing different A_{Vth} parameters in both cases. These results show that V_{th} variability associated to W and L transistor dimensions is affected by defects position in the device channel and other sources of variability such as RDF, LER or LWR. Thus, $\sigma(V_{th})$ can be decoupled to separately account for the W and L dependences in ultra-scaled technologies [152–154].

Another important metric for circuit design and simulation is the possible correlation between critical transistor parameters, like V_{th} and μ. In order to account for the μ of the tested devices, the Y-Function method has been utilized. The Y-Function extraction methodology was originally developed to avoid the dependence of the extracted V_{th} value on μ_{eff} degradation and parasitic series resistance. The method proposes that the $I_{DS}/\sqrt{gm}$ behaves

as a linear function of V_{GS} bias, whose interception with the V_{GS} axis will equal the V_{th} and the slope of the linear fitting results in the μ transistor parameter [148].

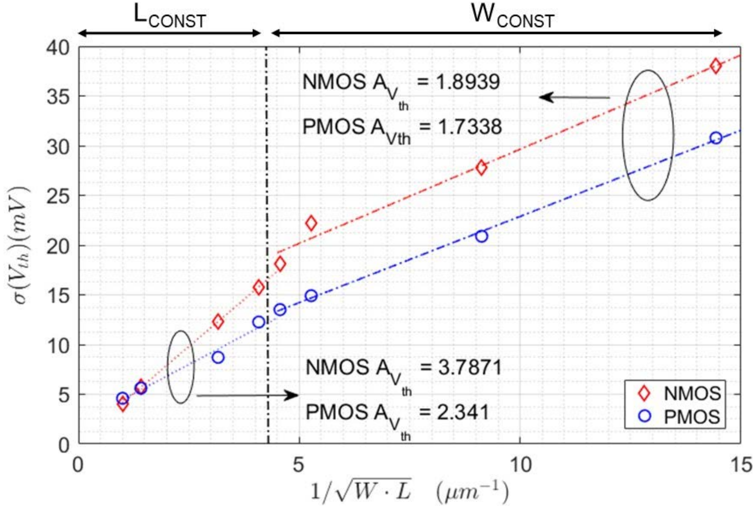


Figure 4.4: Experimental σV_{th} for nMOS (diamonds) and pMOS (circles) transistors as a function of $1/\sqrt{W \cdot L}$. The lines correspond to linear fittings showing different slopes when fixing constant width and length parameters.

The Pearson's correlation between the non-degraded V_{th} and μ parameters are shown in Figure 4.2 (a) and (b) for the smallest, (80nm/60nm = 1.33), and the largest (1000nm/60nm = 16.66) shown in Figure 4.2 (c) and (d) fabricated transistor for both nMOS and pMOS type MOSFETS. Correlation results shown in Figure 4.5 (a) and (b) denotes weak negative correlation, i.e., -0.273 and -0.122 respectively, between V_{th} and μ parameters for the 1.33 transistor ratio. Figures 4.5 (c) and (c), on the contrary, show a larger correlation for the 16.6 transistor ratio, i.e., -0.390 and -0.521 respectively. Finally, Table 4.1 summarizes the statistical TZV data compiled in the present section from device characterization. Moreover, Table 4.1 collects the mean (<>) and the standard deviation (σ) including the total number of tested transistors per geometry. The statistical TZV obtained for the analysis of the 15 ENDURANCE chip samples will be utilized further in this section to obtain the unique transistor degradation under TDV aging phenomena.

Table 4.1: NMOS and PMOS massive Time-Zero variability analysis.

NMOS Geometry (nm)	80/60	200/60	600/60	800/60	1000/60	1000/100	1000/500	1000/1000
# devices	17888	208	208	208	468	468	468	468
$< V_{th} > (V)$	0.3399	0.3918	0.4074	0.4130	0.4144	0.4041	0.3651	0.3504
$\sigma(V_{th})(mV)$	38.03	27.76	22.17	18.09	15.76	12.39	5.727	4.076
$< \mu > (cm^2/V \cdot s) \cdot 10^{-3}$	6.566	9.257	15.79	18.41	20.86	18.08	8.941	6.436
$\sigma(\mu)(cm^2/V \cdot s) \cdot 10^{-3}$	0.4661	0.3999	0.4179	0.4496	0.4287	0.2971	0.06773	0.04116
$Correlation$	-0.273	-0.326	-0.255	-0.394	-0.390	-0.376	-0.130	-0.035
PMOS Geometry (nm)	80/60	200/60	600/60	800/60	1000/60	1000/100	1000/500	1000/1000
# devices	20640	240	240	240	540	540	540	540
$< V_{th} > (V)$	0.4263	0.4325	0.4537	0.4582	0.4625	0.4340	0.3727	0.3588
$\sigma(V_{th})(mV)$	30.77	20.89	14.91	13.50	12.26	8.709	5.613	4.593
$< \mu > (cm^2/V \cdot s) \cdot 10^{-3}$	4.458	6.546	10.50	11.89	13.16	10.98	5.608	4.042
$\sigma(\mu)(cm^2/V \cdot s) \cdot 10^{-3}$	1.004	0.6221	0.2683	0.2282	0.2689	0.1655	0.0485	0.0319
$Correlation$	-0.122	-0.031	-0.352	-0.445	-0.551	-0.521	-0.082	-0.028

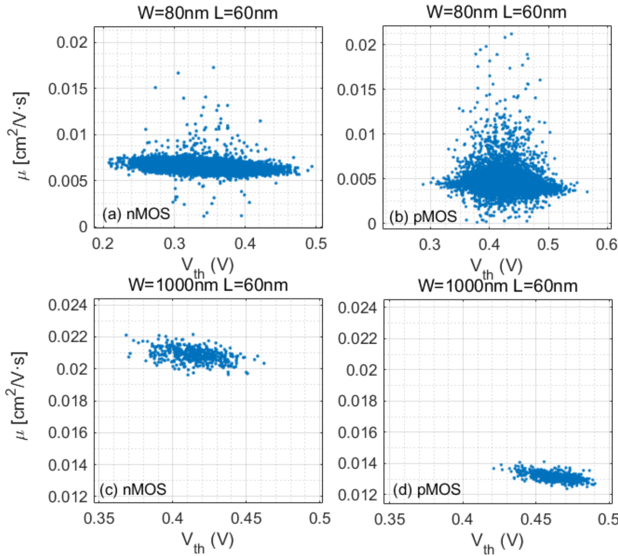


Figure 4.5: Pearson correlation between non-degraded V_{th} and μ electrical parameters for 80nm/60nm nMOS (a) and pMOS (b) and for 1000nm/60nm nMOS in (c) and pMOS (d).

4.1.2 Random Telegraph Noise characterization

As introduced in Section 1.2.1 in Chapter 1, the subject of Random Telegraph Noise (RTN) has become an important concern as a source of variability in deeply scaled CMOS technologies. Significant performance degradation and/or variations can occur in devices and circuits such as SRAMs, ring oscillators or in RRAM, where its significance and impact are expected to increase in newer technology nodes [155, 156]. In this sense, a statistical characterization of the RTN phenomena is required to understand the physics behind it. To do so, hundreds of transistors must be characterized under different biasing and temperature conditions to analyze its behaviour and characterize the time constants associated to the charge and discharge of device defects responsible for the RTN phenomenon.

Figure 4.6 illustrates an example of RTN phenomenon showing a set of 10 different I_{DS}-t current traces of individual pMOS devices with W/L = 80nm/60nm in a time window of 50s utilizing a sampling rate of 2ms. As shown the Figure 4.6, different RTN signals with numerous I_{DS} levels

and widely distributed emission and capture time constants can be observed altering the I_{DS} trace under constant $V_{GS} = 0.6V$ and $V_{DS} = 0.1V$ biasing over time.

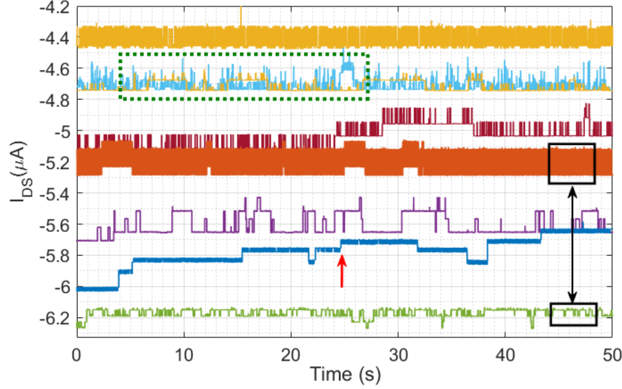


Figure 4.6: A set of 8 different I_{DS} signals characterized under constant biasing conditions showing RTN phenomena behaviour. The red arrow marks a trace with 6 I_{DS} levels while the green dashed box illustrates noisy RTN signals.

As shown in Figure 4.6, the RTN signals of different devices display very distinct features. For instance, devices may display a different number of detectable RTN current levels, e.g., the blue signal shown in Figure 4.6 (pointed out with a red arrow), shows six different I_{DS} levels. Also the amplitude of the current jumps (ΔI_{DS}) caused by different defects can have different magnitudes, like for instance the ΔI_{DS} amplitude difference between the orange trace (ΔI_{DS} = 177nA) and the green trace (ΔI_{DS} = 38nA) denoted with black squares in Figure 4.6. Furthermore, the emission and capture time constants of defects may expand across several decades [33], depending on the particular device defects physics, like for instance, the orange trace in Figure 4.6 that shows fast RTN transitions while the purple RTN signal reveals much slower transitions. This wide variety of RTN signals shown expose that the extraction of the I_{DS} current levels and the ΔI_{DS} values associated to the RTN phenomenon can become a tedious and complicated task.

An RTN study has been executed to statistically observe the behaviour of the defects present in a set of 500 devices with increasing V_{GS} biasing conditions [48]. For this test, five V_{GS} voltages (i.e., 0.5V, 0.6, 0.7V, 1.0V and 1.2V) have been applied sequentially to each transistor from an ENDURANCE

IC chip, keeping a fixed V_{DS} of 0.1V in all cases. Devices have been tested serially by means of the TARS software (see Chapter 3) utilizing a measurement time window of 50s resulting in a total test time of ≈35h.

The objective of the described tests is to obtain the ΔI_{DS} between two I_{DS} levels caused by an RTN transition. To this aim, a three step process has been executed for each of the 2,500 measured traces: first, the background noise has been removed utilizing the Maximum Likelihood Estimation method (MLE), developed in the IMSE-CNM (Sevilla) in the frame of the collaborative project where this thesis is involved [157]. This method digitizes the I_{DS} RTN traces removing the background noise and determining the number and position of the I_{DS} levels present in the trace. The second step consists in determining the ΔI_{DS} values (associated to defects in the device) by means of a full-custom algorithm that is executed in the core of the MLE methodology based on the previously defined I_{DS} levels.

Figure 4.7 shows the empirical CDFs of the encountered ΔI_{DS} for each V_{GS}/V_{DS} biasing condition. As shown in Figure 4.7, the CDFs behaviour indicates that, when the V_{GS} is increased, the ΔI_{DS} increase with a clear positive shift. Nevertheless, for the V_{GS} voltages of 1.1V and 1.2V CDFs, the CDFs are almost overlapped, meaning that changing V_{GS} between 1.1V and 1.2V does not cause a significant variation in terms of ΔI_{DS}.

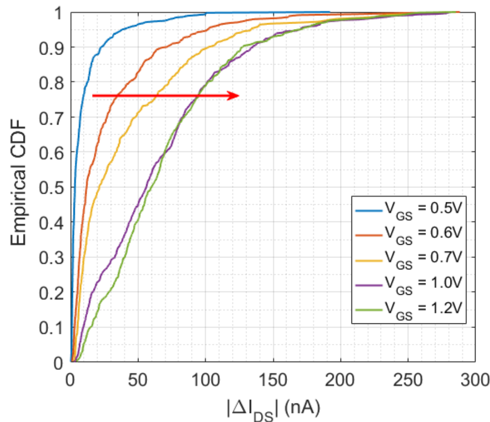


Figure 4.7: Empirical CDFs of the ΔI_{DS} values located in a massive RTN tests in transistors with W = 80nm and L = 60nm dimensions at five different V_{GS} measure voltages with a fixed V_{DS} of 100mV. Red arrows show the tendency of the CDFs for increasing V_{GS} voltage.

As the core of this thesis is focused in device characterization and reliability analysis under BTI and HCI aging phenomena, the results shown in this section for the RTN analysis will not be used for the study of the lifetime prediction.

4.1.3 Bias Temperature Instability characterization

The present section will describe the statistical characterization of the BTI aging phenomena to study reliability effects utilizing the ΔV_{th} degradation metric. Performing statistical characterization utilizing over-than-nominal operation voltages permits to deepen into the degradation effects at the nominal operation voltage to predict, for instance, the expected lifetime of the tested devices as a function of the non-functional devices for a certain ΔV_{th} degradation criterion. With the aim of significantly reducing the total characterization time of accelerated statistical BTI tests, the eSM technique has been utilized by means of the ADPSPM methodology (see Section 3.1 of Chapter 3). The tested DUTs utilized for all the BTI measurements are 80nm/60nm SiON insulator transistors from different samples of the ENDURANCE IC chip. The characterization procedure executed to each transistor under BTI accelerated stress has consisted in four steps:

1) In the first place, before the execution of any accelerated degradation, each device has been initially characterized by the execution of $I_{DS} - V_{GS}$ curve to obtain the non-degraded intrinsic DUT parameters to further compute device shifts due to accelerated stress.

2) In second place, an accelerated BTI tests consisting in five stress-measurement phases has been executed, repeated utilizing sets of non-degraded devices for five V_{GS}/V_{DS} stress biasing conditions as listed in Table 4.2: the $V_{GS} = 1.2V$ the nominal operation voltage of the CORE transistors, and 1.5V, 2.0V and 2.5V over-than-nominal voltages maintaining $V_{DS} = 0V$ during stress. Moreover, the stress times have been increased exponentially starting with a stress time of 1s (i.e., 1s, 10s, 100s, 1,000s and 10,000s) and all recovery phases have been characterized with a $V_{GS} = 0.6V$ and a $V_{DS} = 0.1V$ to monitor de I_{DS}-t behaviour in a measurement window of 100s. In addition, Table 4.2 lists the total number of tested transistors for both NBTI and PBTI tests conducted at $25°C$ and $80°C$ for each stress condition together with a test time comparison between the serial and ADPSPM characterisation methodologies.

In summary, the total amount of time required to characterize 4,268 transistors using serial device characterization, would last for approximately 1 year and 7 months of continuous characterization, resulting in a non-feasible testing time. On the contrary, as shown in Table 4.2, thanks to the ADPSPM methodology, the total test time required is reduced to only ≈25 days. a significant time reduction that allows massive device characterization.

Table 4.2: NBTI and PBTI aging tests numbers description.

		V_{GS};V_{DS} stress conditions			
Test	Temp	1.2V; 0V	1.5V; 0V	2.0V; 0V	2.5V; 0V
NBTI	25°C	500	400	400	400
NBTI	80°C	196	396	396	396
PBTI	25°C	100	100	100	100
PBTI	80°C	196	196	196	196
Total devices		992	1092	1092	1092
Serial		133.3d	146.7d	146.7d	146.7d
ADPSPM		5.7d	6.4d	6.4d	6.4d

Negative Bias Temperature Instability

The NBTI aging phenomena characterization executed in pMOS transistors, Figure 4.8 shows several experimental recovery traces captured after the stress phase of an NBTI aging tests. As shown in Figure 4.8, the equivalent ΔV_{th} shift of the characterized $I_{DS} - t$ trace after each stress phase is shown by means of the non-degraded I_{DS}–V_{GS} curve characterization [60]. All the presented recovery traces show discrete ΔV_{th} discharges that can be visually detected together with the presence of background noise and an important contribution of RTN phenomena that exhibits some of the traces.

The addition of RTN during BTI recovery, demonstrates that BTI and RTN are caused by the same kind of defects. Moreover, previous BTI stress can create new RTN defects that can remain or disappear in a wide range of timescales [57, 158]. The presence of RTN can difficult the analysis of the recovery traces masking the ΔV_{th} BTI-related discharges due to the contribution of different RTN amplitudes and time constants. Thus, Figure 4.9 (a) shows a zoom of the recovery traces of Figure 4.8 to evidence that the presence of a wide variety of RTN signals can appear during BTI-recovery measurements. Additionally, Two zooms are shown in Figure 4.9(b) and (c)

with particular (τ_c, τ_e) time constants and ΔV_{th} amplitudes. To overcome this issue, an RTN removal methodology will be presented in Section 4.2.2 of this chapter to erase RTN fluctuations for a precise BTI behaviour analysis.

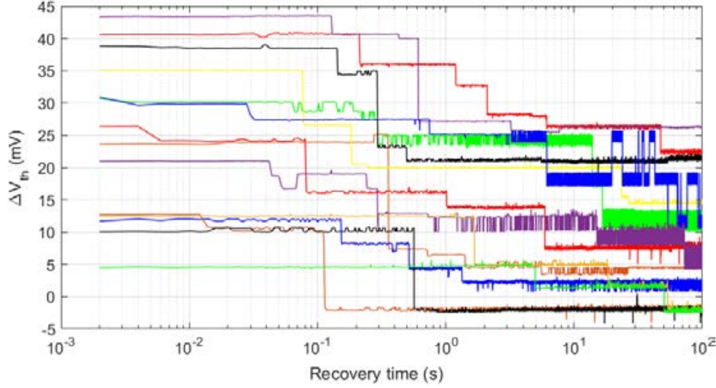


Figure 4.8: NBTI recovery traces showing discrete ΔV_{th} discharges mixed with RTN noise in a logarithmic time scale.

Following with the data analysis, Figure 4.10 shows the $< \Delta V_{th} >$ degradation behaviour as a function of the recovery time window after each accumulated stress time. The ΔV_{th} values recorded at 10 different recovery times (i.e., t_r = 2ms, 5ms, 10ms, 50ms, 100ms, 1s, 10s, 50s, 70s and 100s) have been extracted from all recovery traces for each accumulated stress time (i.e., t_s = 1s, 11s, 111, 1,111, 11,111s). Finally, the $< \Delta V_{th} >$ values are obtained from all the recovery traces at each recovery time for each accumulated stress time. As shown in Figure 4.10, the recovery behaviour during the measurement window shows a clear $< \Delta V_{th} >$ decrease as the recovery time increases due to the discharge of trapped defects during previous BTI stress. Moreover, with increasing accumulated stress time, the $< \Delta V_{th} >$ degradation also increases in all cases as expected. It is important to notice that at the end of the measurement window, the $< \Delta V_{th} >$ does not recover completely, i.e., $< \Delta V_{th} >= 0V$, meaning that large recovery time can show larger $< \Delta V_{th} >$ recovery until the permanent BTI component appears ($\tau_e >>$) measurement window.

With the purpose of obtaining the expected lifetime prediction of the tested devices under NBTI phenomena, the very first recovery point recorded immediately after each BTI stress phase at a recovery time ($t_r = 2ms$) will be utilized. For that purpose, Figure 4.11 shows the empirical CDF of the ΔV_{th}

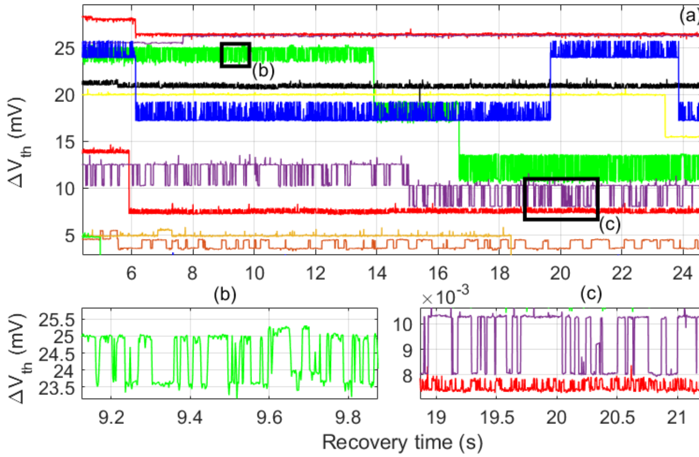


Figure 4.9: Presence of RTN in NBTI recovery traces in (a), while (b) and (c) show the tow zooms in the RTN signals with different amplitudes and τ_c, τ_e time constants.

vales at t = 2ms after each cumulative stress time for each stress voltage. It is important to notice that for 1.2V and 1.5V (Figure 4.11(a) and (b)), which are considered not very high stress voltages, many negative ΔV_{th} appear in the CDFs. This behaviour occurs due the very low degradation obtained from low stress voltages and short accumulated stress times like, for instance the $V_{GS} = 1.2V$ after 1s or 10s of stress. In this case, almost the 30% of the CDFs corresponds to negative ΔV_{th} values.

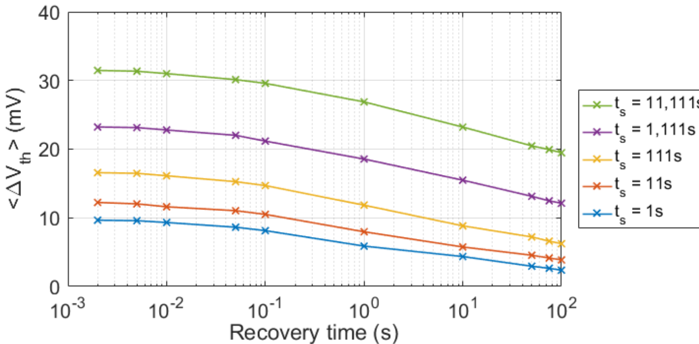


Figure 4.10: $< \Delta V_{th} >$ vs. recovery time for the NBTI test with $V_{GS} = 2.5V$ and $V_{DS} = 0V$.

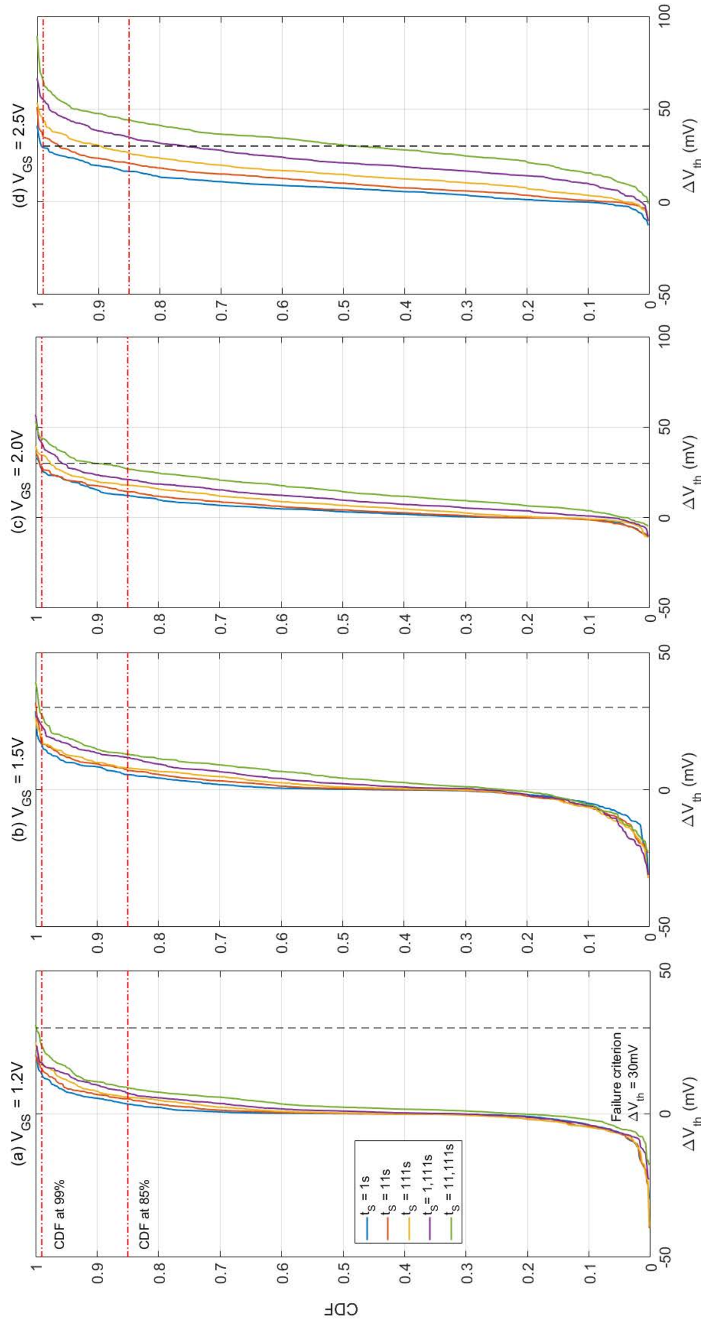


Figure 4.11: NBTI CDFs of the first ΔV_{th} recovery point after each stress phase for the four stress voltages.

To accurately evaluate the lifetime in terms of percentage (%) of non-functional devices, the CDFs reported in Figure 4.11 are evaluated at different % levels. In order to obtain the lifetime extrapolation at each evaluated CDF %, the five ΔV_{th} values obtained are plotted in a log-log plot, where the Y-axis corresponds to the $log_{10}(\Delta V_{th})$ in volts and the X-axis stands for the accumulated stress time at each stress voltage in seconds. The next step consists in executing an power law extrapolation using the five ΔV_{th} points for each stress voltage. Figure 4.12(a) and (b) show two examples where the five ΔV_{th} degradation values plotted as a function of the accumulated stress time for the 85% in (a) and the 99% in (b) of the CDF % levels. Moreover, a black dashed line determines the location of the ΔV_{th} degradation criterion of 30mV used in this study as the device non-functional frontier shown in Figure 4.12 with red dashed lines while the ΔV_{th} = 30mV are depicted with black dashed lines.

To obtain the lifetime prediction at the ΔV_{th} degradation criterion, a power law extrapolation utilizing the five ΔV_{th} values is executed for each stress voltage at each evaluated CDF % level as shown in the two examples of Figure 4.12(a) and (b) as coloured straight lines. Then, the extrapolated lifetime values are obtained, as the intersection between the power law extrapolation with ΔV_{th} degradation criterion. This intersection reveals the required stress time to obtain a ΔV_{th} of 30mV for each CDF % and represents that all devices below the CDF % are considered functional, i.e., because they have not reached the ΔV_{th} degradation criterion, while the rest of the devices are considered non-functional. In this sense, Figure 4.13 depicts the expected lifetime extrapolation in seconds for the non-functional devices for each BTI stress voltage at the temperature of $25°C$.

Several observations can be extracted from the expected lifetime results shown in Figure 4.13: first, due to the significantly low degradation achieved for low stress voltages and/or low stress times, the last 40% of non-functional devices, i.e., from 60% to 100%, cannot be used for the lifetime prediction and thus, Figure 4.13 only shows up to the 60 % of the non-functional devices. Second, the lifetime prediction shown for the NBTI stress of 1.2V and 1.5V have been extrapolated utilizing the lifetime results from 2.0V and 2.5V stress voltages because of the non-conclusive results achieved due to the low degradation.

As shown in Figure 4.13, for high BTI stress voltages, i.e., V_{GS} = 2.5V, the expected lifetime is significantly reduced where the 60% of devices are considered non-functional before $\approx$ 3 hours. Nevertheless, the expected

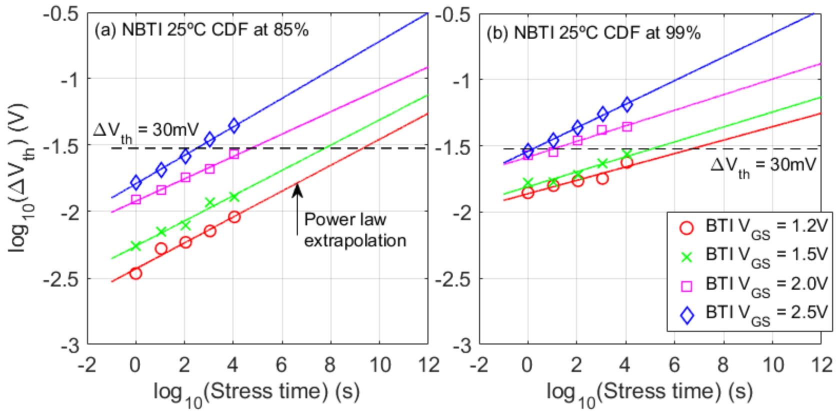


Figure 4.12: Lifetime extrapolation procedure for a ΔV_{th} criterion of 30mV.

lifetime prediction at the 1.2V nominal operation condition reveals a greater device functionally margin. In this sense, as indicated in Figure 4.13 by a red and green dashed lines, at 1.2V the 8% of the tested devices will be considered not functional in 1 year, while the ≈ 28% of the devices will fail in 10 years under solely NBTI aging phenomena at 25°C.

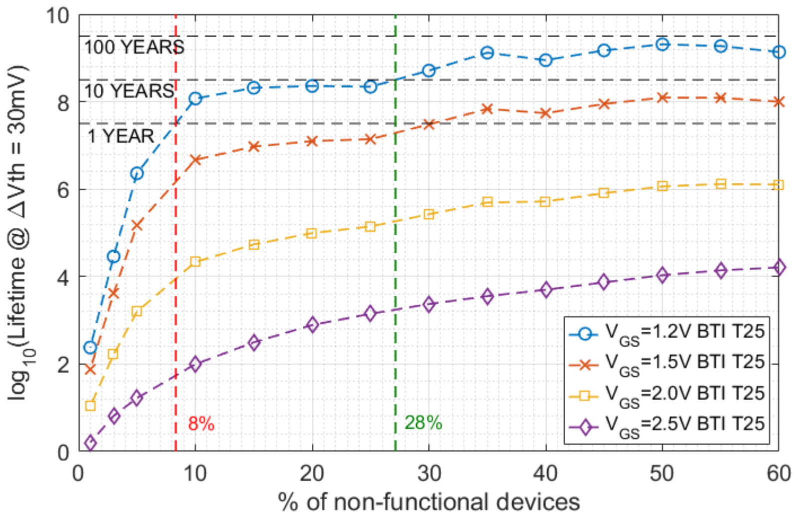


Figure 4.13: Expected lifetime vs. the percentage of non-functional devices for NBTI at 25°C.

An important consideration when dealing with BTI degradation is its dependence with temperature conditions. In this sense, and as listed in Table 4.2, a set of NBTI experiments have been executed at 80°C to obtain the expected lifetime under elevated temperature conditions. The temperature of 80°C is chosen because commercial products, e.g., commercial processors and GPUs, work in a temperature range between 50°C and 100°C thus, the experimental high temperature has been chosen inside the typical range of commercial products. In order to evaluate the temperature effects of the NBTI degradation the same analysis conducted at 25°C has been executed for the NBTI measurements conducted at 80°C. In this sense, Figure 4.14 shows an illustrative comparison of the $< \Delta V_{th} >$ as a function of the accumulated stress time at the recovery time of t_r=2ms and t_r = 100s of all tested devices comparing NBTI at 25°C and 80°C for a stress voltage of $V_{GS} = 2.5V$ and $V_{DS} = 0V$.

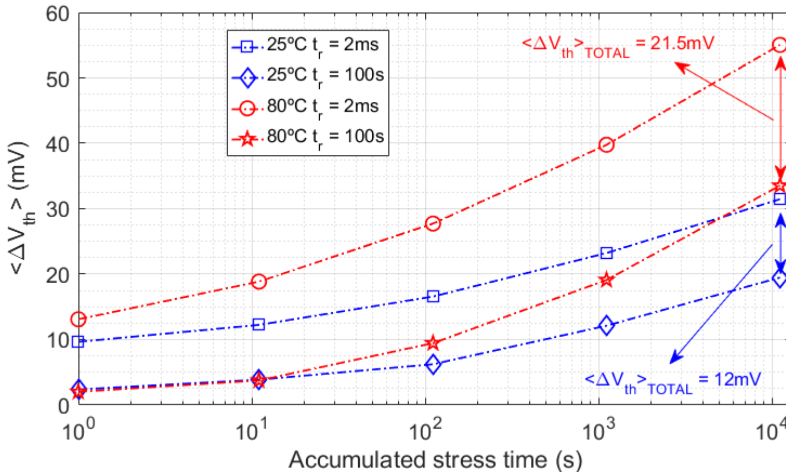


Figure 4.14: Comparison the $< \Delta V_{th} >$ at t_r = 2ms and t_r = 100s recovery data samples after 25°C and 80°C NBTI stress voltage of $V_{GS} = 2.5V$ and $V_{DS} = 0V$.

As shown in Figure 4.14, the ΔV_{th} values at t_r = 2ms after high temperature stress (red circles) are always higher than the ones at lower temperature (blue squared) as expected. Moreover, a visual comparison of the total recovery at each accumulated stress time for both temperatures is always higher for 80°C (i.e., the gap between circles and stars in red) than for 25°C (i.e., the difference between squares and diamonds in blue) in all cases. For instance, for

the accumulated stress time of 11,1111s Figure 4.14 shows a total $< \Delta V_{th} >$ recovery achieved at 80°C of 21.5mV in contrast to the 12mV of total recovery achieved at 25°C. These results demonstrate that as higher the stress/recovery temperatures, higher degradation/recovery achieved is always higher when compared with low temperatures. These two behaviours can be explained because at higher temperatures, NBTI stress captures more defects than at 25°C and also creates new interface states that are charged as well. Moreover, τ_c and τ_e are also changed due to high temperature stress in agreement with Section 1.3.1 in Chapter 1.

In terms of NBTI temperature degradation comparison, Figure 4.15 shows the expected lifetime comparison between NBTI at $25°C$ and $80°C$ at the nominal operation voltage of 1.2V. The results show that the expected lifetime under $80°C$ drastically decreases at each shown % of non-functional devices. For instance, if a conservative 20% of non-functional devices is imposed as a design frontier, as shown in Figure 4.15 with a black dashed vertical line, at $25°C$, the expected device lifetime is $\approx$ 10 years. Nevertheless, at $80°C$ the lifetime is drastically reduced to $\approx$ 21 days. On the other hand, if a wider percentage is chosen, for instance a 60%, the expected lifetime increases for $25°C$ close to 100 years while for $80°C$ gets close to 1 year. In this scenario, the final application will impose the required percentage of non-functional devices to estimate the lifetime of the devices in a particular technology as well as the particular characterization conditions.

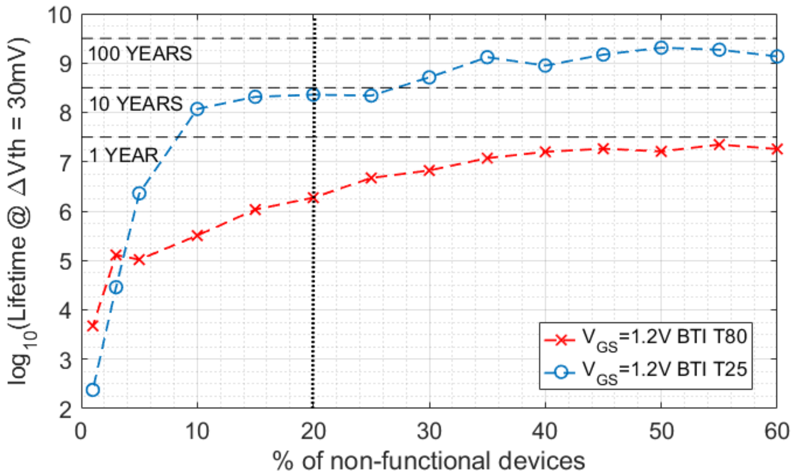


Figure 4.15: Comparison between BTI degradation at $25°C$ and $80°C$.

Positive Bias Temperature Instability

As have been detailed in Table 4.2, statistical measurements have been conducted in order to account for the expected lifetime prediction also for PBTI aging phenomena. In this sense, Figure 4.16 (a) and (b) shows two CDFs where the ΔV_{th} obtained at $t_r = 2$ms after each stress time for the stress voltages of $V_{GS} = 2.5V$ in (a) at $25°C$ and $80°C$ in (b). From the data shown in both CDFs, two important observations can be extracted: first, the ΔV_{th} values plotted in Figure 4.16 (a) shows that the stress of 1s produces higher degradation than all the accumulated stress which is not expected. On the contrary, in Figure 4.16 (b) the CDFs are sorted as the stress time increases as expected. Nevertheless, what both figures have in common is the very low degradation is achieved by the PBTI stress when compared with the NBTI degradation as shown in Figure 4.11 (d) at $25°C$ and Figure 4.14 at $80°C$, respectively.

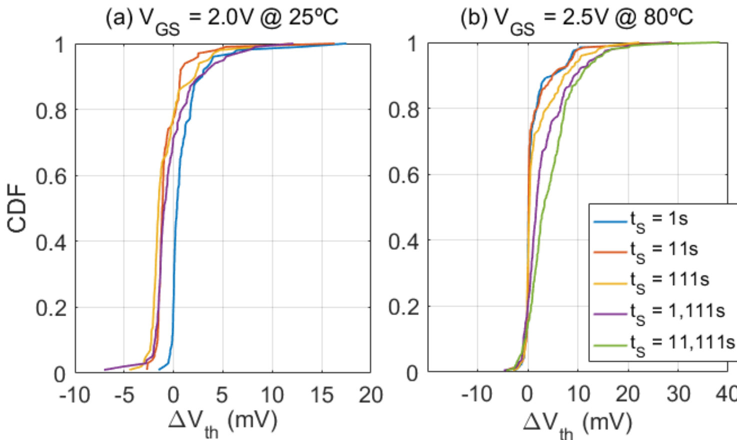


Figure 4.16: PBTI CDFs of the first recovery point after each stress phase at (a) PBTI with stress $V_{GS} = 2.0V$ and (b) PBTI with stress $V_{GS} = 2.5V$

A clear example of the low ΔV_{th} degradation achieved under PBTI stress of $V_{GS} = 2.5V$ and $V_{DS} = 0V$ is shown in Figure 4.17 for $25°C$ in (a) and $80°C$ in (b). The $< \Delta V_{th} >$ degradation reveals very low degradation is achieved when compared with the NBTI results (see Figure 4.10). The data obtained from PBTI Figure 4.16 and Figure 4.17, clarifies that the degradation, in the nMOS devices of the ENDURANCE IC, is not sufficiently relevant and will not be used for the extraction of the expected lifetime prediction.

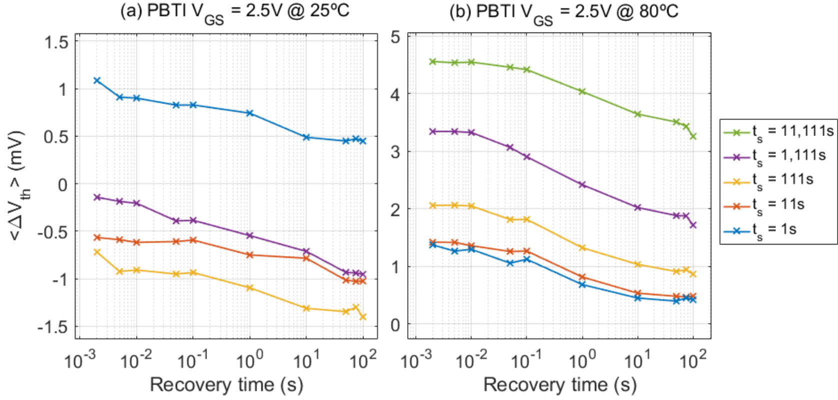


Figure 4.17: $< \Delta V_{th} >$ degradation as a function of the recovery time at $V_{GS} = 2.5V$ for (a) $25°C$ and $80°C$ after each accumulated stress.

4.1.4 Hot Carrier Injection characterization

As described in Chapter 1, HCI aging phenomena has become an important source of device and circuit variability in nowadays advanced technology nodes. In this sense, to study the effects of the HCI degradation, in both, nMOS and pMOS devices, a statistical analysis will be conducted in this section to obtain the expected lifetime prediction under different stress biasing conditions to later compare HCI against BTI degradation at the nominal operation voltages.

For the statistical characterization of the HCI aging effects, Table 4.3 lists the number of tested nMOS and pMOS devices with W = 80nm and L=60nm. During both nMOS and pMOS HCI experiments, the same eSM sequence utilized for the BTI analysis has been utilized for the HCI tests. Moreover, the RTN affectation at the recovery measurements is lower for the HCI case than for the BTI. For ultra-scaled MOSFET transistors the worst-case accelerated stress conditions occurs when $V_{GS} = V_{DS}$ [78, 159, 160]. Thus, for the HCI tests conducted in this section with $V_{GS} = V_{DS}$ for 1.2V, 1.5V, 2.0V and 2.5V stress conditions for nMOS and pMOS MOSFETs as shown in Table 4.3 while all recovery phases have been executed utilizing $V_{GS} = 0.6V$ and V_{DS} = 0.1V.

For the HCI experiments described in Table 4.3, the total DUT number together with a time comparison between serial and the novel ADPSPM

Table 4.3: Number of devices utilized in for the HCI experiments.

	V_{GS}, V_{DS}			
Test	1.2V, 1.2V	1.5V, 1.5V	2.0V, 2.0V	2.5V, 2.5V
HCI NMOS	50	250	200	210
HCI PBTI	498	392	150	400
Total devices	548	642	350	610
Serial	67.8d	57.05d	23.66d	62.42
ADPSPM	3.114d	3.426d	1.794d	3.107d

characterization technique are listed. For the four stress voltage conditions used for HCI testing, a total of 2,150 transistors have been characterized requiring only 11.4 days of testing. On the contrary, if conventional serial device characterization would have been used, the required laboratory test time would have lasted for approximately 7 months of continuous laboratory characterization.

As described for the BTI aging phenomena study, the same procedure has been utilized to statistically obtain the HCI expected lifetime as a function of the % of non-functional devices. In this sense, Figure 4.18 shows the CDFs of ΔV_{th} values obtained immediately after each accumulated stress time for the four $V_{GS} = V_{DS}$ stress conditions. A clear difference between the HCI and BTI CDFs can be observed: the ΔV_{th} degradation experienced by the tested devices is higher for the HCI case, shown in Figure 4.18, than the results shown for NBTI in Figure 4.11.

Figure 4.19 shows the $< \Delta V_{th} >$ as a function of the recovery time under HCI stress voltage of $|V_{DS}| = |V_{GS}| = 2.5V$ for nMOS devices in Figure (a) and for pMOS devices in Figure (b). As shown in both figures, during the recovery window of 100s, the $< \Delta V_{th} >$ degradation suffered by the nMOS devices in (a) behaves as a permanent damage. On the contrary, the HCI stress applied to the pMOS transistors in (b) shows a small $< \Delta V_{th} >$ recovery trend after a ≈100ms time for 1s, 11s and 111s of HCI stress. Nevertheless, the small $< \Delta V_{th} >$ recovery observed practically disappears after the 1,111s and 11,111s of accumulated stress and the damage in nMOS devices is higher than in pMOS.

What is also interesting to observe from the experimental results of Figure 4.19, is that for the HCI stress in pMOS devices, the entire ΔV_{th} damage after 1s, 11s and 111s of stress is higher than the ΔV_{th} produced in the nMOS ones for the same accumulated stress time. However, the overall damage

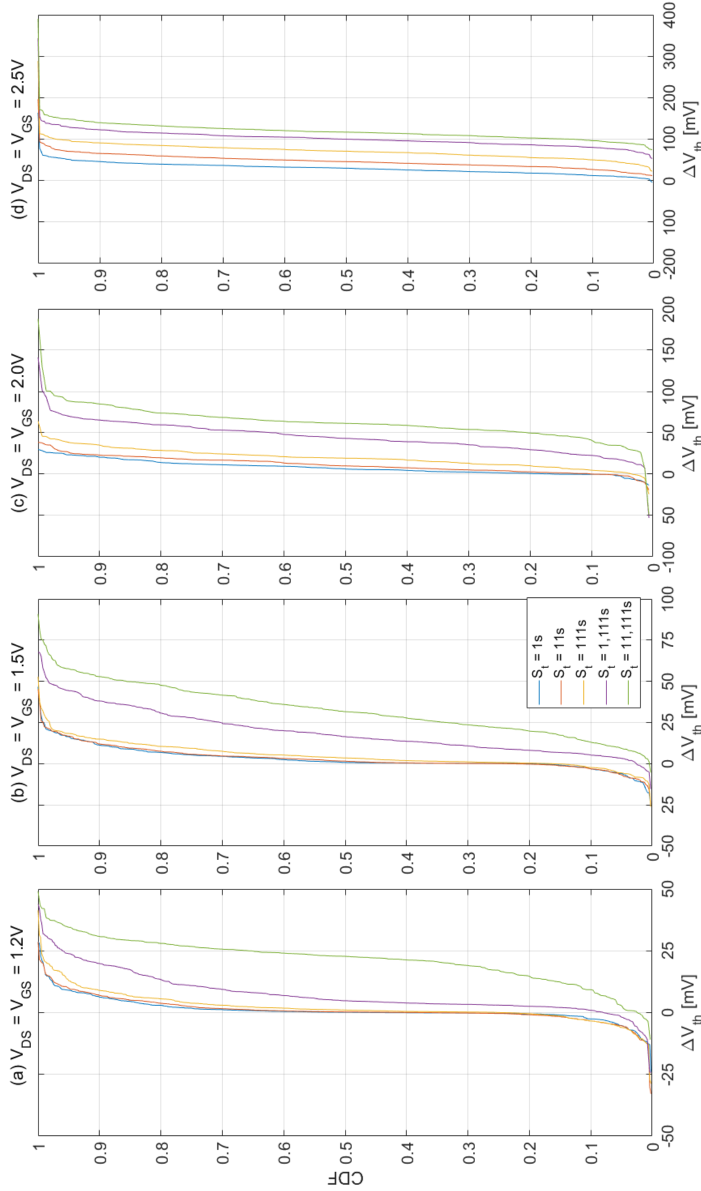


Figure 4.18: HCI in pMOS transistors: CDFs of the first ΔV_{th} recovery point after each stress phase for the four stress voltages.

after 1,111s and 11,111s of stress is higher at the nMOS transistors than in the pMOS ones, in agreement with the HCI literature [78]. With the results shown so far for HCI degradation, Figure 4.20 depicts the expected lifetime prediction for the HCI phenomena for the nMOS devices in (a) and for the pMOS devices in (b) as a function of the percentage of non-functional devices at the temperature of 25°C.

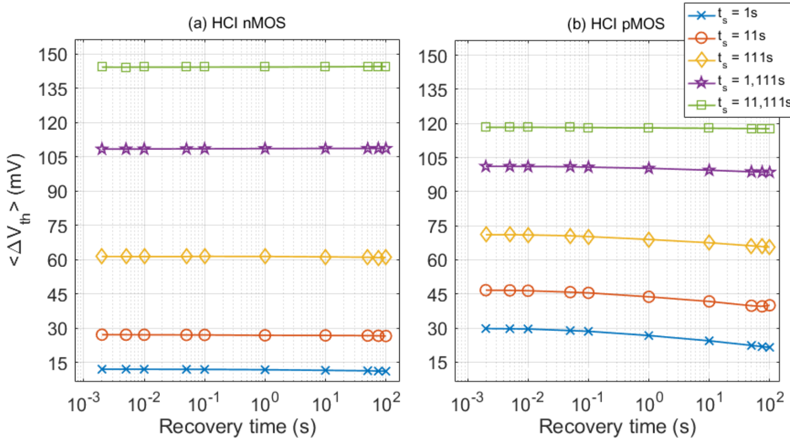


Figure 4.19: $< \Delta V_{th} >$ vs. recovery time for the HCI aging results at the stress voltage of $V_{GS} = V_{DS} = 2.5V$ for nMOS transistors in (a) and pMOS transistors in (b).

Due to noise and non-conclusive results obtained at the HCI stress voltage of $V_{GS} = V_{DS} = 1.2V$, the lifetime prediction shown in both nMOS and pMOS graphs in Figure 4.20 have been extrapolated utilizing the results obtained from the $V_{GS} = 1.5$V, 2.0V and 2.5V and $V_{DS} = 0$V stress conditions. Moreover, as a consequence of the negative values of ΔV_{th} extracted from both nMOS and pMOS HCI tests (see the pMOS CDFs shown in Figure 4.18), the percentage of non-functional devices is shown up to the 85% for the nMOS case and up to the 75% in the pMOS case, as can be seen in Figure 4.20(a) and (b) respectively.

The expected lifetime shown in Figure 4.20 reveals that the damage suffered by the tested 80nm/60nm devices is slightly more severe under HCI in nMOS than in pMOS transistors. This degradation behaviour is in agreement with the TZV results shown in Figure 4.5 (a) and (b) in Section 4.1.1 where, the mobility of many pMOS devices is equal or higher than the nMOS ones. For

instance, at the operation voltage of 1.2V the expected lifetime for nMOS and pMOS devices show approximately the same expected lifetime (around ≈ 300 hours). None of the reported lifetime results at 1.2V reaches the lifetime expectation of 1 year for the defined degradation criterion of ΔV_{th} = 30mV. In the following section, a comparison between the analyzed BTI and HCI will be shown.

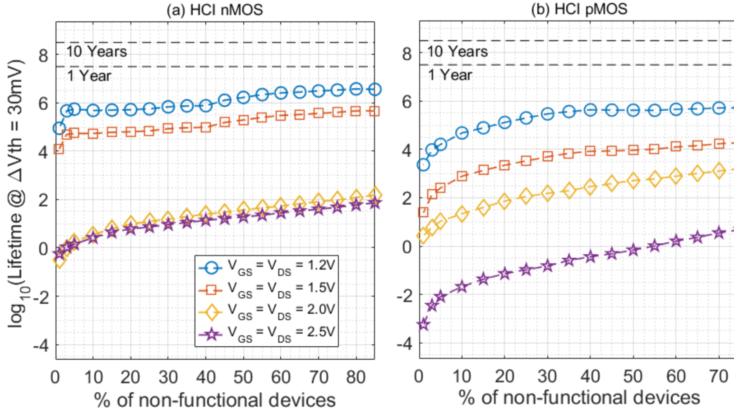


Figure 4.20: Expected lifetime vs. the % of non-functional devices for HCI stress at 25°C for nMOS transistors in (a) and pMOS transistors in (b).

4.1.5 TDV lifetime comparison

During accelerated device stress under BTI and HCI phenomena, degradation mechanisms rapidly degrade reliability of the tested devices. In this sense, Figure 4.21 depicts a comparison between NBTI, HCI in nMOS and pMOS $< \Delta V_{th} >$ degradation as a function of the accumulated stress from 1s up to 11,111s. The first observation that can be extracted from Figure 4.21 is that increasing the stress time increases the $< \Delta V_{th} >$ shift for the three studied phenomena showing that the HCI damage aggravates the degradation more than BTI.

For the NBTI case, the absolute recovery between the first (t_r = 2ms) and last (t_r = 100s) values of recovery remains constant even though the accumulated stress time increases (see Figure 4.14). On the contrary, the HCI aging reveals two different behaviours: for the nMOS devices a significantly

low recovery is observed because no gap exists between the t_r = 2ms and t_r = 100s recovery values (i.e., difference between red squares and diamonds in Figure 4.21). On the contrary, in the pMOS HCI case a decreasing $|\Delta V_{th}|$ recovery can be observed while increasing the accumulated stress time (i.e., difference between blue circles and stars). Moreover, the HCI damage in pMOS devices is higher than for the nMOS ones before the accumulated stress time of 111s as not expected. But this trend changes between 111s and 1,111s accumulated stress where the damage suffered by the nMOS transistors becomes more severe than the pMOS transistors as expected. It is also interesting to observe how the degradation imposed by HCI aging (i.e., blue and red dashed lines) is significantly larger than the BTI (i.e., green dashed lines) case for long accumulated stress times.

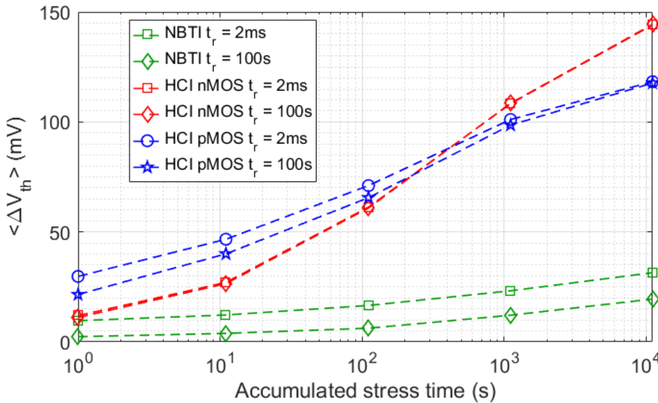


Figure 4.21: $< \Delta V_{th} >$ degradation behaviour between t_r = 2ms and t_r = 100s recovery data values during 100s of recovery window for NBTI and HCI phenomena as a function of the accumulated stress time. All aging experiments have been conducted utilizing $V_{GS} = 2.5V$ and $V_{DS} = 0V$ stress and $V_{GS} = 0.6V$ and $V_{DS} = 0.1V$ recovery biasing conditions.

At this point, and with the statistical results presented so far for BTI and HCI aging phenomena, Figure 4.22 shows a comparative graph where the expected lifetime prediction is presented. Moreover, the figure includes dashed lines that determine 1, 10 and 100 years of lifetime frontier. The traces presented in Figure 4.22, reveal that the less harmful degradation phenomena correspond to the NBTI at 25°C, in this case defining a conservative barrier of non-functional devices percentage equal to 20% (marked in Figure 4.22 with a vertical red dashed line), the expected lifetime prediction of almost

10 years is obtained. Nevertheless, the expected lifetime results for NBTI at 80°C and HCI for nMOS and pMOS at 25°C, show a very low expected lifetime prediction of less than 25 days.

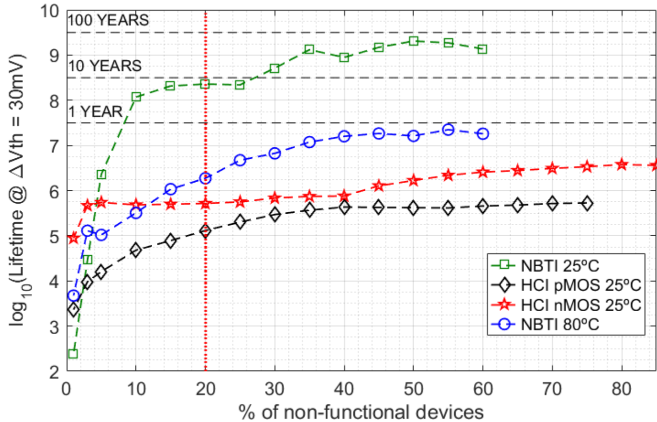


Figure 4.22: Comparison graph of the expected lifetime prediction as a function of the % of non-functional devices between NBTI at 25°C and 80°C and HCI for nMOS and pMOS devices. PBTI degradation is not present in this comparative graph due its significantly low achieved degradation.

The results presented in this section can be utilized to apply design guard bands during the analog or digital design state to increase the expected lifetime of the fabricated devices and circuits. Moreover, depending on the specific application or circuit to be designed, all measurements and accelerated stress strategies can be modified to accommodate and reproduce more reliable results for the final circuit functionality to obtain more accurate lifetime predictions based on particular circuits.

4.2 Extraction of MOSFET parameters for circuit reliability simulation

During BTI/HCI device aging testing, a precise $I_{DS} - t$ to ΔV_{th} conversion is required to accurately monitor the degradation after the application of electrical stress. Device parameter drifts must be extracted from the measured data, for which first/second-order transistor parameters extraction methods

can be found in the literature [149, 161]. However, the accuracy of these methods is usually object of controversy. Moreover, these methods are not focused on the transistor compact models (e.g., the BSIM4 MOSFET model) that are eventually used for circuit simulations [162, 163].

From a circuit design point of view, the device aging-induced degradation should be mapped into those transistor model parameters to accurately reproduce the experimentally measured loss of reliability. In this sense, the present section will describe the parameters extraction methodologies developed in this thesis for an accurate transistor reliability analysis. Parts of these section have been already published in [164–167].

4.2.1 A model parameter extraction methodology including time-dependent variability

Accurate extraction of transistor parameters affected by accelerated TDV tests, like threshold voltage (V_{th}) or mobility (μ), has become a key step when using electrical transistor parameters from laboratory measurements, such as I_{DS}–V_{GS} or I_{DS}–t for its further utilization in circuit simulations. In this scenario, the present section, will describe two novel methodologies intended for the extraction of VTH0 and U0 BSIM4 model card parameters and a methodology to convert I_{DS}–t measurement traces from BTI/HCI aging tests in to the equivalent $\Delta VTH0$–t model card parameter.

To do so, the Mean Squared Error (MSE) metric will be utilized to compare experimental measurements with a set of HSPICE simulated I_{DS}–V_{GS} traces. In this scenario, depending on the type of data to be analyzed (i.e., I_{DS}–V_{GS} or I_{DS}–t curves) this section present the two different methodologies:

1. The VTH0-U0-BSIM extraction method, is utilized to obtain VTH0 (threshold voltage of the short channel device at zero substrate bias) and U0 (low-field mobility) BSIM4 model card parameters from experimental I_{DS}–V_{GS} curves measured on both fresh and aged devices.
2. The accurate I_{DS}–t - to - $\Delta VTH0$ conversion method is utilized to evaluate the shift of VTH0 BSIM parameter after the application of accelerated stress from experimental I_{DS}–t traces. The method translates the measured I_{DS}–t into the equivalent $\Delta VTH0$, including BTI/HCI aging-related and RTN phenomena on the $\Delta VTH0$ compact model parameter.

In order to visually verify the accuracy of BSIM 4 model card parameters extraction methodology, Figure 4.23 (a) and (b) shows an example of the best match achieved between an experimental I_{DS}–V_{GS} curve of a pMOS MOSFET transistor with W=80-nm/L=60-nm and the HSPICE simulated I_{DS}–V_{GS} curve in blue with the lowest MSE. As clearly shown in the Figure, the obtained I_{DS}–V_{GS} HSPICE simulated curve with the minimum MSE perfectly matches with the experimental curve since the simulated and the experimental curves are overlapped.

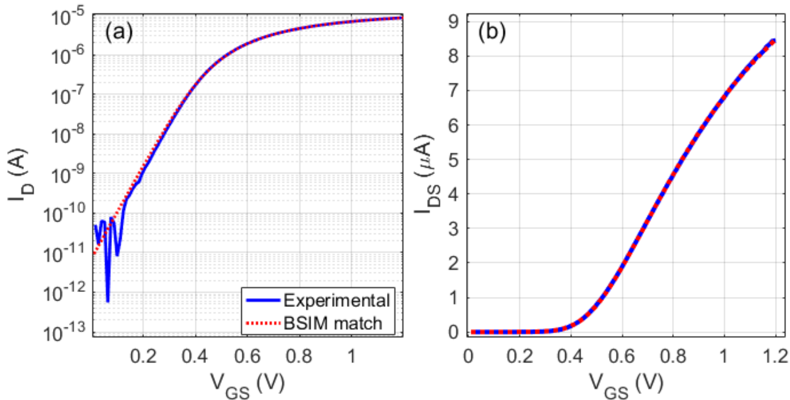


Figure 4.23: Matching between an experimental $I_{DS} - V_{GS}$ curve in blue and the HSPICE simulated $I_{DS} - V_{GS}$ curve in red with the lower MSE between 100,000 simulations with VTH0/U0 uniformly distributed pairs.

To quantitatively account for the MSE of a large set of experimentally measured I_{DS}–V_{GS} curves, Figure 4.24 shows the MSE between simulated and experimental curves. Slightly high MSE can be found in the histogram of Figure 4.24 due to the contribution of noise for low V_{GS} voltage biasing, as well as, the presence of RTN present in some the traces.

NBTI and HCI tests has been conducted at room temperature on 1,200 pMOS and 600 nMOS transistors (both with W/L=80nm/60nm), respectively. To evaluate the transistor parameter extraction and the I_{DS}–t - to - $\Delta VTH0$ conversion methodologies. In this sense, 4 BTI tests (each one involving 300 CMOS transistors) and 6 HCI tests (each one involving 100 transistors) have been carried using the ADPSPM technique (see Section 3.1 of Chapter 3), with a 6-cycle eSM scheme. During the measurement phases, the I_{DS}–t curves are registered at $V_{GS} \approx V_{th}$ and $V_{DS} = 100mV$ during 100s followed

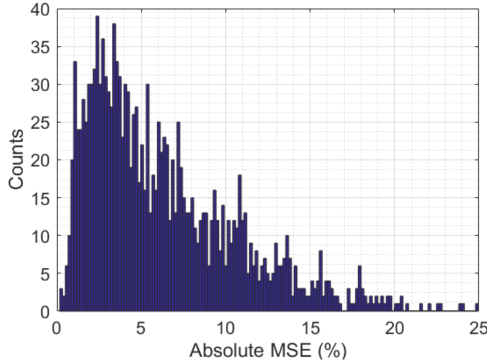


Figure 4.24: Histograms of the MSE obtained between 1,050 experimental $I_{DS} - V_{GS}$ curves and the simulated HSPICE curves with minimum MSE.

by an I_{DS}–V_{GS} curve for further analysis of the transistor parameters shifts. By means of the ADPDPM technique executed by the TARS software, the total aging test time required to test 1800 devices with 6-SM cycles, has been reduced from 6.37 years of continuous serial testing to only ≈12.5 days. The two proposed methodologies are detailed below:

A) **The VTH0-U0-BSIM methodology for model card parameters extraction:** this extraction method consists in finding the best match, i.e., minimum Mean Squared Error (MSE), between the I_{DS}–V_{GS} experimental curve and a set of 100,000 HSPICE simulated I_{DS}–V_{GS} curves. The VTH0/U0 sample pairs were selected with the foundry provided TZV BSIM v4 compact model fully covering fresh and degraded $I_{DS} - V_{GS}$ scenarios at 25°C. At this point, to obtain the model card parameters of the experimental curves, the best matching curve between the experimental and a HSPICE simulated curve that gives the minimum MSE is selected, and the VTH0 and U0 parameters are extracted from the HSPICE simulation parameters. The obtained model card parameters can be later included in the circuit simulator to observe the impact of degradation on the circuit performances.

With all the statistical VTH0 and U0 extracted model card parameters, a study of the BSIM parameters correlations is shown in Figure 4.25. Figures (a) and (b) show correlations between the $\Delta U0$ and $\Delta VTH0$ after 1s in (a) and after 1.000s in (b) of HCI stress. High uncorrelated degradation can be found between $\Delta VTH0$ and $\Delta U0$ for the largest stress voltages. Correlation between $\Delta VTH0$ and $\Delta U0$ can be seen when the stress voltages applied are kept below the maximum operation technology voltages.

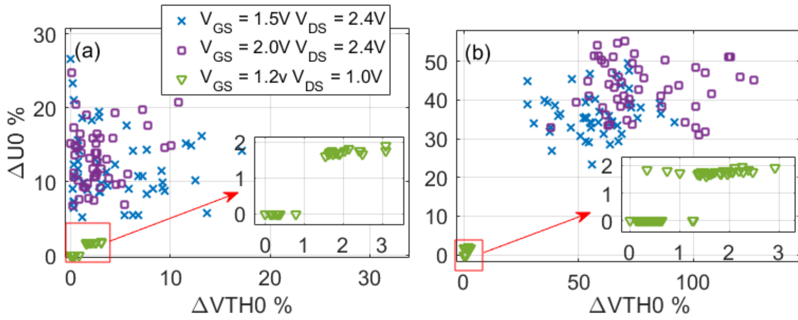


Figure 4.25: $\Delta U0$ as a function of $\Delta VTH0$ obtained from the $I_{DS} - V_{GS}$ experimental curves after 1s (a) and 1000s (b) of HCI stress.

In Figure 4.26 (a) and (b) the correlation between $\Delta U0$ and $\Delta VTH0$ after BTI and HCI aging tests are shown, respectively. Figure 4.26 (a) shows a significant correlation between $\Delta U0$ and $\Delta VTH0$ parameters after 1s of BTI stress can be found but, as the stress time increases the correlation drastically decreases. Figure 4.26(b) shows a high correlation between $\Delta U0$ and $\Delta VTH0$ after HCI stress if V_{GS} and V_{DS} voltages are kept below 1.5V but for $V_{DS} \leq$ 1.5Vs the correlation decreases.

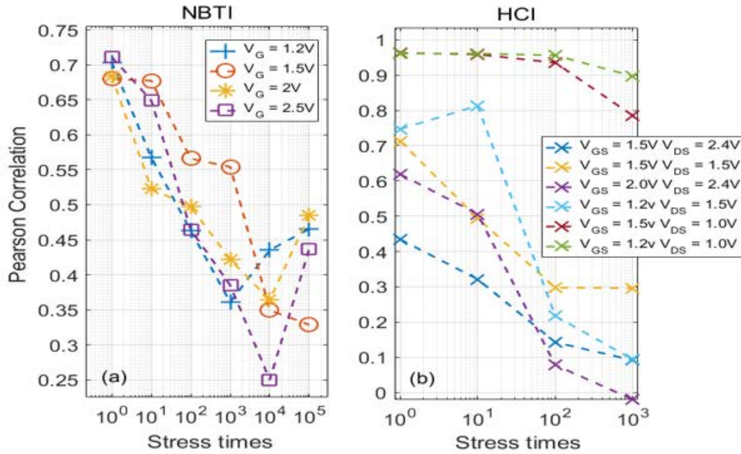
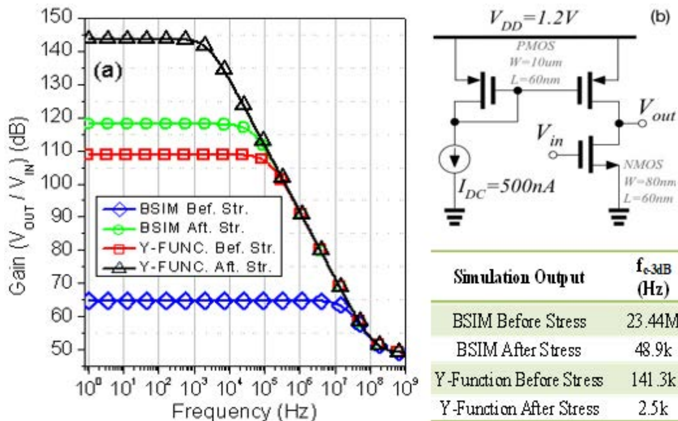


Figure 4.26: Pearson correlations between $\Delta U0$ and $\Delta VTH0$ computed from the experimental $I_{DS} - V_{GS}$ curves obtained for (a) 300 (80nm/60nm) pMOS devices under BTI stress and (b) 100 (80nm/60nm) nMOS devices under HCI stress were considered for each one of the stress conditions shown.

For the comparison between the Y-Function and the BSIM parameter extraction methods, a straightforward simulation of an analog amplifier is carried out and the results are shown in Figure 4.27, by replacing the foundry provided VTH0 and U0 parameters of its transistors by those obtained on the degraded ones using the VTH0-U0-BSIM parameter extraction method. The results obtained from both simulations demonstrate that the use of the parameters provided by the Y-Function extraction method, compared to the proposed method, may result, for instance, into false-positive end-of-life predictions [168]. Using a 10kHz cut-off frequency (fc-3dB) as lifetime metric (i.e., the amplifier is considered non-functional for a fc-3dB below 10kHz), after a HCI stress of 1.000s at V_{DS} = 2.0V, V_{GS} = 2.4V, with the parameters obtained with the proposed method fc-3dB = 48.9kHz, while when obtained with the Y-Function method a circuit malfunction results with a fc-3dB = 2.5kHz, a wrong prediction of the amplifier lifetime.



Simulation Output	$f_{c\text{-}3dB}$ (Hz)
BSIM Before Stress	23.44M
BSIM After Stress	48.9k
Y-Function Before Stress	141.3k
Y-Function After Stress	2.5k

Figure 4.27: (a) Simulation results of the gain of a common source analog amplifier (b), using VTH0/U0 model card parameters of fresh and HCI-degraded devices extracted with the proposed method and the Y-Function method. The Y-Function method extracted parameters that resulted in an underestimation of 2 orders of magnitude of the real cut-off frequency before and after the HCI test compared with the proposed parameter extraction method.

B) **The I_{DS}–t - to - $\Delta VTH0$ conversion methodology:** this method allows studying the mean deviations of the VTH0 and U0 parameters of BTI and HCI stress. By means of the generated HSPICE 100,000 I_{DS}–V_{GS} curves, the I_{DS}-t curve, registered during the measurement phase (V_{DS} =

$V_{GS} = 1.2V$), is converted into the equivalent $\Delta VTH0$–t trace, assuming that the possible degradation of U0 parameter (if exists) does not significantly influence the $\Delta VTH0$ change [169]. The new I_{DS}–t - to - $\Delta VTH0$ conversion methodology, results in a $\Delta VTH0$–t trace shifts due to BTI- and/or RTN-like behaviour in terms of model card parameters variations in a three-step process:

1. Calculation of VTH0 and U0 from post-stress I_{DS}–V_{GS} curves: the algorithm extracts the VTH0 and the U0 non-degraded parameters from the experimental I_{DS}–V_{GS} curve registered after the i-th eSM cycle.
2. Selection of suitable simulated I_{DS}–V_{GS} curves: the algorithm discards those HSPICE simulated I_{DS}–V_{GS} curves with different U0 parameter from the one obtained from step 1. When doing this, it is assumed that the change of I_{DS} is mostly related to changes in VTH0. The method also discards those I_{DS}–V_{GS} curves with a lower U0 value than that obtained from the I_{DS}–V_{GS} curve measured in the (i-1)-th cycle because it is assumed that the U0 parameter does not recover significantly.
3. I_{DS}–t conversion to $\Delta VTH0$–t: to obtain the equivalent $\Delta VTH0$–t trace, the algorithm searches in the remaining set of simulated I_{DS}–V_{GS} curves the minimum I_{DS} difference at the V_{GS} and V_{DS} voltages applied during the measurement phase for each I_{DS}–t sample. Then, the $\Delta VTH0$ is obtained from the difference of the non-stressed VTH0 and the one obtained for each I_{DS}–t data sample.

To evaluate the proposed I_{DS}–t - to - $\Delta VTH0$ conversion methodology, 6 different I_{DS}–t curves after 1000s of BTI stress are shown in Figure 4.28 (a), and their resulting $\Delta VTH0$ waveforms are shown in Figure 4.28 (b). Figure 4.28 (a) reveals rising current jumps, i.e., interface trap detrapping, in the I_{DS}–t curves, which are observed as falling $\Delta VTH0$ jumps in Figure 4.28 (b). Background noise and RTN are also mapped in the $\Delta VTH0$–t trace.

From the analysis of the statistical data obtained from the ENDURANCE chip by means of the TARS software and the utilization of the ADPSPM technique, the accurate VTH0-U0-BSIM extraction methodologies presented in this chapter have been used to capture the aging-induced degradation in the VTH0 and U0 BSIM4 model card parameters. Moreover, a novel I_{DS}–t - to - $\Delta VTH0$ conversion method that can be useful for circuit designers to improve lifetime prediction and reliability-aware circuit design. In the next section a methodology to remove the presence of the RTN phenomena will be presented to avoid RTN-masking issues when localizing ΔV_{th} discharges in BTI and HCI aging recovery traces.

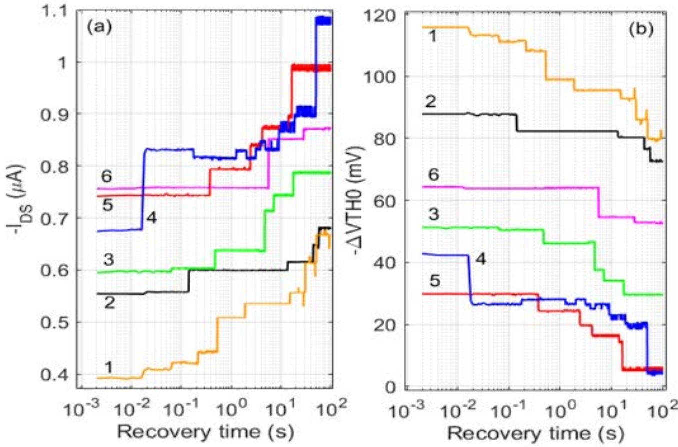


Figure 4.28: Conversion of the $I_{DS} - t$ curves to $\Delta VTH0 - t$ measured on 6 pMOS devices after BTI stress. (a) Current traces after stress with $V_{GS} = 2.5V$ and $V_{DS} = 0V$. The figure shows the ΔI_{DS} current with current increments mixed with background noise and RTN phenomena. (b) Equivalent $\Delta VTH0 - t$ waveforms obtained using the proposed ΔI_{DS}–to–$\Delta VTH0$ conversion method.

4.2.2 A smart noise- and RTN-removal methodology for transistor parameter extraction

With nowadays technology scaling trend imposed to advanced technology nodes, accurate characterization time-dependent variability effects related to the trapping/detrapping in/from device defects, like for instance RTN, BTI and HCI, have re-emerged as a serios reliability thread in the IC design. Moreover, simulation and implementation of reliability-aware techniques under TDV, have become a key step in today's nanometric technologies design and simulation phases [170]. In order to do so, appropriate TDV compact models, like the Probabilistic Defect Occupancy (PDO) model [98], are essential tools utilized to account for and clearly distinguish the BTI-related 'slow' defect discharges, from the RTN-related 'fast' defects switching behaviour, which is in fact a not straightforward procedure.

In this scenario, this section describes a novel and smart methodology of transistor defect parameters characterization utilized to remove the presence of RTN and background noise from BTI or HCI aging measurements traces after

the application of accelerated stress. The methodology, evaluates the emission τ_e and the induced threshold voltage fluctuation (η) of solely individual BTI-related slow defect discharges, found during the measurement of BTI degradation after stress. The proposed methodology follows the flow diagram shown in Figure 4.29 to accurately extract the defect discharges parameters described hereinafter:

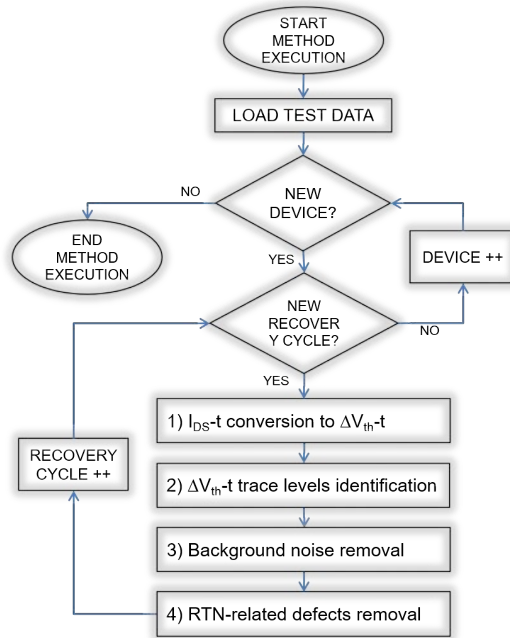


Figure 4.29: Diagram flow for the extraction method of τ_e, η.

Description of the method: the proposed defect parameters extraction methodology utilizing as input data massively characterized BTI/HCI aging related measurement traces. In this particular case, this chapter will focus in accelerated BTI tests carried out by means of the ADPSPM method based on the execution of the eSM technique. For each device involved in a BTI aging test, the method treats each BTI recovery trace individually. First, the I_{DS}-t recovery trace recorded after the stress is converted into the equivalent ΔV_{th}-t trace. Second, by means of the WTLP method [112, 155] (see Section 1.4.2 in Chapter 1), all ΔV_{th} levels of the $\Delta V_{th}-t$ trace are obtained extracted from the trace. Figure 4.30 (a), shows an example of a BTI recovery trace plotted in

terms of ΔV_{th}-t, where significant contributions of background noise and RTN can be easily observed in the inset in Figure 4.30 (a) and a particular trace zoom between t=20s and t=45s in Figure 4.30 (b), respectively. Finally the RTN and the background noise are removed from the trace to only account for the aging-related behaviour. The described m 5-step procedure is described in detail in the following:

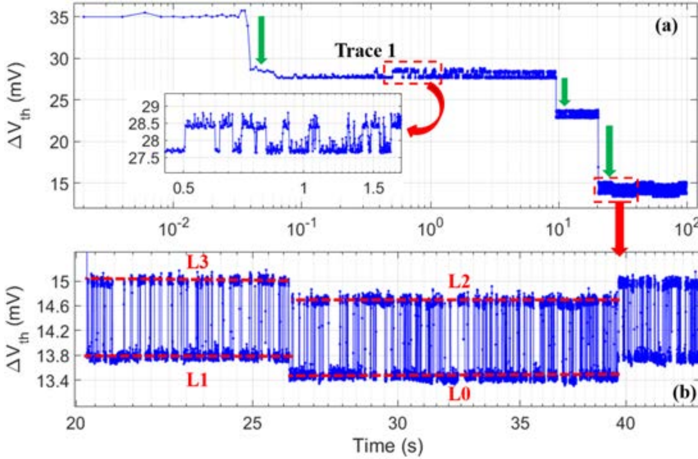


Figure 4.30: (a) Example of a ΔV_{th} trace constructed from the measured I_{DS} recovery data. In (b) a zoom-in of (a) showing the combination of two fast defects and background noise on top of a ΔV_{th} level.

1) I_{DS}-t to ΔV_{th}-t conversion: for each tested device, the I_{DS}-t recovery traces are converted into the equivalent ΔV_{th}-t trace by means of the non-degraded $I_{DS} - V_{GS}$ curve characteristic [60]. For instance, as shown in the equivalent ΔV_{th}-t trace of Figure 4.30 (a), three charge emissions (denoted with green arrows) occur during the recovery time. Figure 4.30 (b) shows a zoom in the recovery trace of (a) that reveals fast capture/emission transitions, switching between 4 different ΔV_{th} levels (i.e., L0 to L3), the joint contribution of two individual RTN signals, and mixed with background noise: one switching between ΔV_{th} levels L0-L1 and the other one switching between L1-L3 and between L0-L2.

2) Identification of ΔV_{th}-t trace levels: the next step consists in the application of the WTLP method to each recovery trace to identify the number and magnitude of the ΔV_{th} levels present in the trace. For instance, Figure

4.31 (a) shows the WTLP resulting from the analysis of trace 1 in Figure 4.30 (a). Analyzing the diagonal of the WTLP, four groups of populated data regions, separated in the figure by green dashed arrows, can be distinguished. Each one of the data groups corresponds to a different ΔV_{th} level present in the recovery trace that can be easily related with the different levels (RTN- and BTI-like transitions) present in Figure 4.31 (b) highlighted with red arrows.

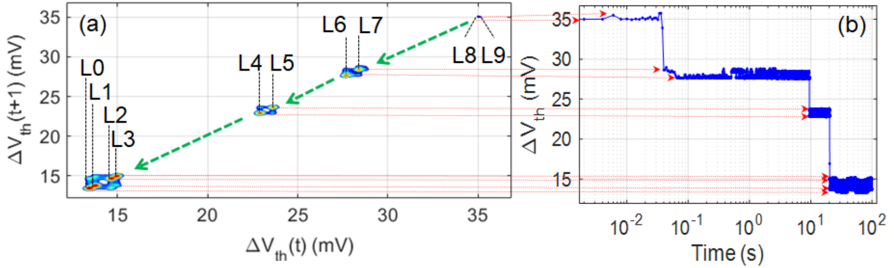


Figure 4.31: (a) WTLP resulting from the analysis of Trace #1 in Figure 4.30(a). (b) Corresponding ΔV_{th} level in the recovery trace for each red-colored zone in the WTLP in (a).

Transitions from one data group to the next one are considered as BTI-related defect emissions at a specific τ_e and the difference between two ΔV_{th} levels correspond to the η of the discharged defect. Fast RTN transitions on top of each ΔV_{th} level can also be clearly distinguished, as red-coloured regions in the diagonal (i.e., in the ΔV_{th} levels L0 to L9) from the less populated regions located perpendicularly to the diagonal (i.e., ΔV_{th} level transitions). Moreover, each ΔV_{th} level of the WTLP is associated with each ΔV_{th} level present in the recovery trace as shown in Figure 4.31 (b) with red arrows.

To clearly distinguish the ΔV_{th} values associated to the RTN transitions detected in Figure 4.30 (b), a zoom-in of the WTLP (red square in Figure 4.31 (a)) is shown in Figure 4.32. The zoom clearly shows the 4 red regions across the diagonal, which correspond to the four different ΔV_{th} levels (i.e., L0, L1, L2 and L3). The blue regions out of the diagonal indicate the transitions between levels (e.g., L1 to L3). For instance, at $t \approx 25s$ the signal switches between levels L1 and L3, and later, at $t \approx 30s$ the signal switches between L0 and L2. These levels result from the joint combination of the two individual RTN signals. The results in Figure 4.32 clearly demonstrate that the application of the WTLP to the ΔV_{th} recovery trace allows an accurate location of all ΔV_{th} levels while distinguishing between RTN and BTI contributions.

3) Background noise removal: once the ΔV_{th} levels have been identified with the WTLP, the method procedure assigns the closest ΔV_{th} level to each sample in the ΔV_{th}-t trace. This step quantizes the ΔV_{th} trace, removing the background noise and keeping only ΔV_{th} levels associated to RTN and BTI. For instance, Figure 4.33 (a)-(c) displays three ΔV_{th} traces (in blue), showing high ΔV_{th} degradation due to stress, together with the ΔV_{th} trace reconstruction (in red) without background noise. During recovery, different ΔV_{th} levels can be observed because of defect emissions mixed with RTN phenomena, as shown in detail in the zoom-in plots in Figure 4.33 (b) and (c).

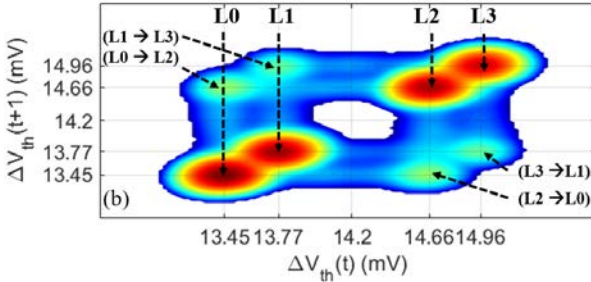


Figure 4.32: Zoom-in of the WTLP in Figure 4.31(a) showing, for the two RTN signals in Figure 4.30(b), the ΔV_{th} levels in the diagonal of the WTLP and the corresponding transitions between ΔV_{th} levels outside the diagonal.

4) Removal of RTN-related transients: in order to distinguish between RTN and BTI contributions, a Transition Matrix (TM), a square matrix that stores the transitions between different ΔV_{th} levels in the recovery trace, is defined. The dimension of the TM is the total number of ΔV_{th} levels found (ten level in this example), where each ΔV_{th} level is denoted as LN where N accounts for the number of levels (e.g., N goes from 0 to 9 in the example). The rows of the TM are defined as the initial ΔV_{th} levels (i.e., i-th ΔV_{th} level), while the columns are defined as the final ones (i.e., (i+1)-th ΔV_{th} level). During the ΔV_{th} trace analysis, three different ΔV_{th} level transitions can be distinguished:

1. Case (i): No change of the ΔV_{th} level. The ΔV_{th} level of the i-th data sample equals the (i+1)-th ΔV_{th} level.
2. Case (ii): Defect charging. The (i+1)-th ΔV_{th} level is larger than the i-th ΔV_{th} level. For instance, samples switching from level L1 to L3.

3. Case (iii): Defect discharging. The (i-th)+1 ΔV_{th} level is smaller than the i-th ΔV_{th} level. For instance, switching from level L3 to L1.

To construct the TM, the ΔV_{th} recovery trace is swept by its (ii, ii+1) elements and the transition location counter of the TM initial-final ΔV_{th} level is incremented when needed. Case-(i) transitions will lay in the main diagonal of the TM while case-(ii) and case-(iii) transitions will be located above and below the TM main diagonal, respectively. For instance, in Trace 1 of Figure 4.33 (a) ten ΔV_{th} levels have been detected and the resulting 10x10 TM is formed with all ΔV_{th} transitions identified. In order to remove fast RTN transitions from the traces, the method analyzes the data in the TM distinguishing the following two cases:

BTI/HCI emission recognition: this type of defect emissions are characterized by a unique defect discharge to a lower ΔV_{th} level. In the TM, this appears as a '1' in the corresponding element below the TM main diagonal, and a '0' in the TM symmetric position. Analyzing the TM in Figure 4.34 (a), three different emissions, without any further capture, can be found marked with blue circles: from initial to final ΔV_{th} levels L4-L3, L6-L4 and L8-L7.

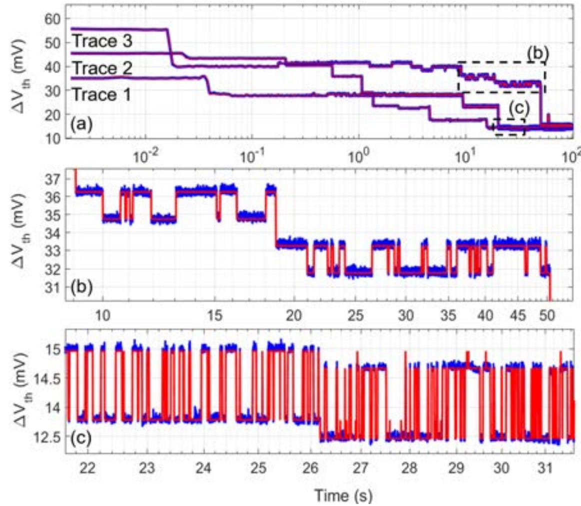


Figure 4.33: (a) Recovery traces (blue) with the reconstructed trace without noise (red). (b), (c) Zoom-in of the traces showing RTN (in blue) and the trace reconstruction without background noise (in red).

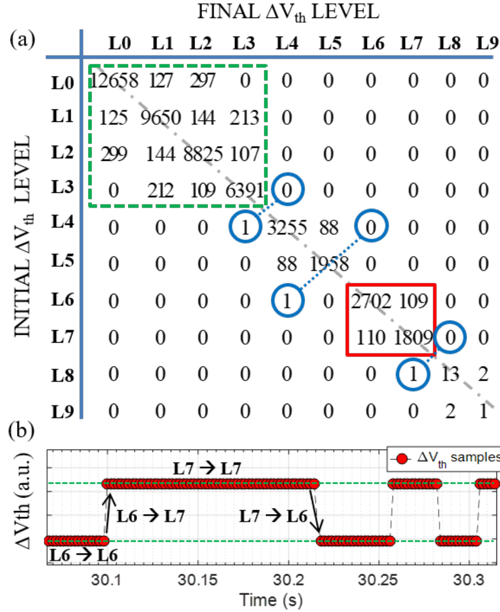


Figure 4.34: (a) TM corresponding to trace #1 in Figure 4.30 (a). The green dashed box defines cases (i), (ii) and (iii) of the joint RTN in (b). The the red solid box defines the transitions of the RTN shown in the inset of Figure 4.30 (a). Blue circles indicate the BTI discharges. (b) Illustrative example of the two ΔV_{th} level transitions in the RTN signal in the inset in 4.30 (a).

RTN recognition: the method identifies multiple and consecutive transitions between two distinct ΔV_{th} levels. For instance, all transitions from/to levels L0 to L3 of the combined RTN signals in Figure 4.30 (b) can be found TM of Figure 4.34 (a) with a green dashed box. Moreover, an example of a single defect RTN transient in the TM is highlighted with a solid red square box, which corresponds to the RTN signal shown in the inset of Figure 4.30 (a). The main diagonal indicates the number of samples at the same ΔV_{th} level (i.e., initial and final ΔV_{th} levels L6 or L7), while 110 defect discharges (i.e., initial L7 to final L6) and 109 defect charges (i.e., initial L6 to final L7) take place in the RTN signal.

5) Defect parameters extraction. the last step is the extraction of the τ_e and η parameters of the slow defect emissions. This is done by locating the elements below the TM main diagonal showing single discharge that have a symmetrical zero above the TM main diagonal. The η value of each detected

defect is obtained by calculating size of the ΔV_{th} drop as shown in Figure 4.35 with the 3 clean recovery traces. To allow the evaluation of the τ_e associated to the defect, when the TM is constructed, also the times at which (ii, ii+1) transitions occur are saved. When a slow discharge is encountered, the τ_e is assigned to the computed η value so the defect is characterized by the tuple (τ_e, η).

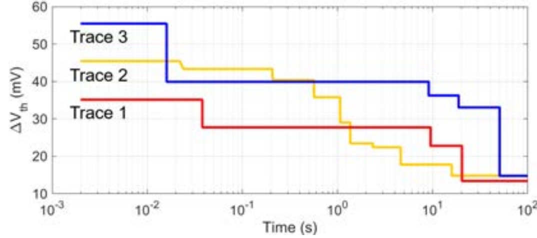


Figure 4.35: Filtered recovery traces without background noise and/or RTN.

In order to evaluate the validity of the described method, four individual BTI experimental tests conducted at room temperature, each one involving 248 CMOS transistors with 8 different channel sizes (see Table 2.1 in Section 2.1.1 of Chapter 2). All devices have been tested using a 6-cycle SM scheme, in which the duration of the stress phases is increased exponentially (i.e., 1s, 10s, 100s, 1000s 10,000s and 100,000s), while the recovery phase time is always 100s. During the stress period, 1.2V, 1.5, 2.0 and 2.5V V_{GS} voltages have been considered, while for measurements $V_{GS} \approx V_{th}$ and $V_{DS} = 100mV$. All tests have been conducted automatically by utilizing the ADPSPM technique reducing the total tests time from ≈3.5 years, using serial characterization to only ≈ 7 days, thanks to the ADPSPM algorithm. From the automated analysis of the data, a total of 10,582 slow emissions have been identified. Figure 4.36 shows the statistical results obtained from the analysis of all (τ_e, η) tuples extracted by using the method. Figure 4.36 (a) shows that $< \eta >$ decreases with transistor area increase.

If the RTN related transients were not removed from the recovery traces before the slow emissions identification, many of 'false' events (with equal ΔV_{th} value as the RTN amplitude) would have been mistakenly considered during the $< \eta >$ calculation. Therefore, the resulting $< \eta >$ value for all tested geometries, would have been closer to the ΔV_{th} of the fastest RTN, masking the actual $< \eta >$ of the BTI-related defects. Moreover, Figure 4.36 (b) shows that η is exponentially distributed. The results agree with those in the literature [50], which supports the validity of the methodology.

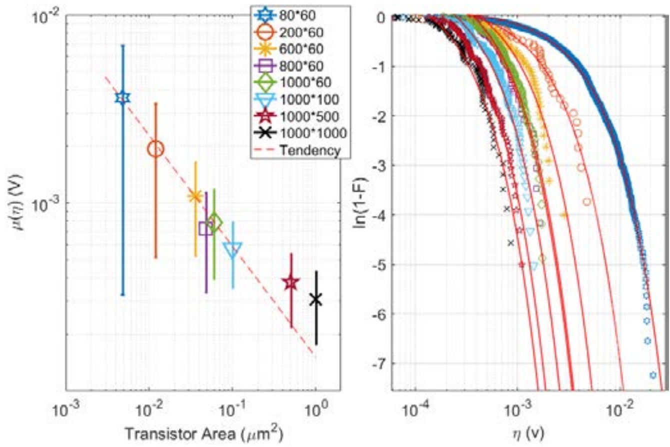


Figure 4.36: (a) Distribution of the $<\eta>$ values of the BTI emissions as a function of the transistor area. (b) Exponential distribution of the η values, with exponential fitting extracted from BTI tests for each geometry.

As shown in Figure 4.35, the described smart noise- and RTN-removal smart method identifies and removes the present RTN and the background noise of the recovery traces to avoid artifices during the detection of BTI/HCI defect discharges. Moreover, from the results shown in Figure 4.36, more than 10,000 defects emissions can be used as input parameters to stochastic aging compact models like the PDO (see Section 1.3.5 in Chapter 1).

The use of stochastic models to simulate large circuits, e.g., CMOS multiplier, under aging conditions, results in highly costly computational resources where simulations can take long times. In this regard, the next section will describe a new GPU-based simulation framework designed to exploit the computational power of modern GPU systems to significantly reduce aging simulation time.

4.2.3 BTI/HCI-aware circuit simulation using a GPU-based approach

Accurate transistor modelling under BTI/HCI phenomena consists in two main challenges: a significant time is required to characterize an statistically representative set of devices to properly calibrate the aging models and also

long times are required to simulate large circuits under TDV phenomena . Both challenges originate from unique uncorrelated BTI behaviour in each device revealing no information about other devices degradation under the same stress condition. Thus, many devices are measured to obtain probability distributions and each device must be modelled and then simulated individually.

Commercial circuit simulator uses deterministic compact BTI models like MOSRA [171] or FTCM [172] that are not suitable to model the stochastic BTI/HCI variability thus, stochastic models like PDO [98] or TTOM [172] must be employed instead. These models are computational complex as they are optimized to fit measured data that results in aging modelling times in the order of minutes per device. Thus, utilizing SPICE circuit simulations are extremely time-consuming resulting in computationally intensive modelling simulations. Therefore, to employ a BTI variability model in circuit simulation, the algorithm must be able to model thousands of MOSFETs devices in a feasible simulation time.

In this scenario, the present section describes a novel GPU-based BTI variability model that employs massive parallelism (beyond 1000 processing cores) found in graphic cards to model thousands of MOSFETs in seconds by simulating devices concurrently. An overview of the entire approach is given in Figure 4.37. Therefore, together with the novel defect parameter extraction methodology presented in Section 4.2.1 of this chapter, accurate model statistical parameters are provided, while the PDO model itself enables large circuit simulations in (>100,000 MOSFETs) in a few minutes.

Background Probabilistic Defect Occupancy Model: by means of the PDO model, each defect $D = (\tau_c, \tau_e, \eta)$ captures/emits carriers, resulting in ΔV_{th} degradation in devices. Each defect is specified by its capture time τ_c, τ_e and the induced ΔV_{th} degradations η as described in Section 1.3.5 in Chapter 1. The PDO evaluates the occupancy probability of defects with particular τ_e and τ_c distributions under arbitrary biasing waveforms (AC or DC) [98]. Moreover, a log-normal bi-variate distribution is utilized for simplicity to describe the defect parameter distribution.

Calibrating the Model: to calibrate the PDO model, the TARS software together with the ADPSPM measurement technique is used to statistically obtain the emission time τ_e and induced degradation η for each defect across all measured devices. Then the MATLAB function ”fitgmdist” is used, which performs a regression fit of bi-variate Gaussian distributions. using the logarithmized values as the defect distribution is a bi-variate log-normal

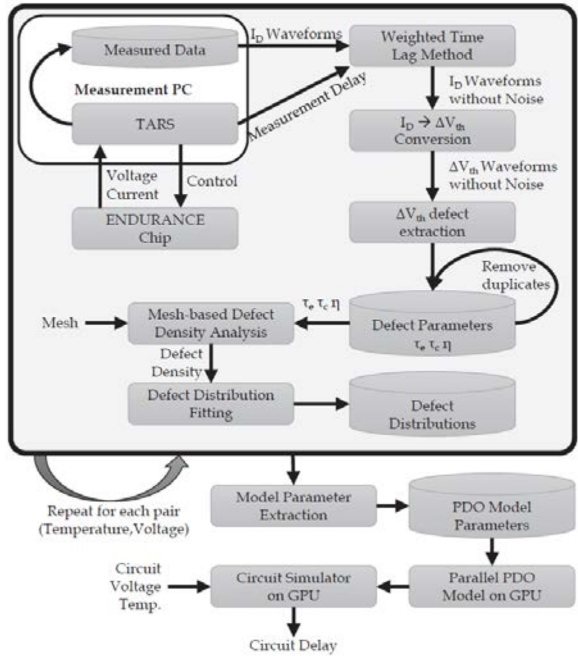


Figure 4.37: Overview of the BTI/HCI variability modelling approach.

distribution [98]. By providing logarithmic data, ”fitgmdist” can fit a bivariate Gaussian distribution with a covariance matrix against our extracted defect parameter data sets. At each voltage and temperature combination, a defect distribution is obtained.

Defect Parameters Distribution: after the aging characterization, the measurements are processed with WTLP (see Section 1.4.2 in Chapter 1) to extract the defect parameters of each defect. Then ”fitgmdist” MATLAB function fits the defect distribution $D(\tau_c, \tau_e)$ to histograms of τ_c, τ_e. The result is a defect distribution shown in Figure 4.38. Then the η histogram is used to fit the exponential distribution of η, as shown in Figure 4.39(a). Figure 4.39 (b) illustrates how regression to the mean ensures that the values become more representative where more transistors are characterized (sample size).

Parallel PDO BTI Model using CUDA programming: simulation times for defect centric BTI models are too large for aging variability. However, graphic cards offer massive computational resources with massive parallelism,

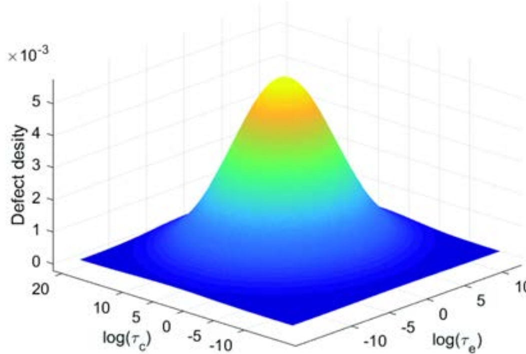


Figure 4.38: D(τ_c, τ_e) defect distribution fit in MATLAB for 100 devices with 194 defects each providing a (τ_c, τ_e) pair. Four MSM cycles with stress V_{GS} of 2.5V and exponential increase stress times, i.e., 1s, 11s, 111s and 1,111s, and a fixed recovery time of 100s are used. The defect distribution verifies long capture times and emissions within 100s of measurement window at room temperature.

for instance 2048 cores in the NVIDIA GTX980. This enables to simulate up to 1024 devices in parallel to decrease simulation times to milliseconds per device as shown in Figure 4.40. To implement PDO on the graphic card, it is programmed in CUDA, a C-variant targeted towards massive parallel programming of NVIDIA graphic cards. Then, the PDO is used to individually model the devices, in which individual defects are randomly occupied or unoccupied according to the occupancy probability P_{occ}. Therefore, each device is modelled as a CUDA block containing up to 72 compute threads where each thread models a single defect.

Parallel PDO Model Performance: figure 4.40 (a) highlights how the parallel implementation is able to model 100,000 devices in less than 120s that results in 1.2ms simulation time per single device while the original MATLAB implementation [98] took about 5 minutes per device. A later version of the PDO model described sequential C programming (shown in orange) already outperforms [173], due to algorithmic optimization. Nevertheless, the massively parallel PDO implementation presented in this section utilizing CUDA on GPU framework enables large scale circuit simulations. To verify this statement, a parallel SPICE simulator has been used to simulate 32 and 64-bit multipliers with 11288 and 42534 devices, respectively, as shown in Figure 4.40 (b). Each device had a unique set of defects, resulting in a unique occupancy state and thus, unique ΔV_{th} degradation behaviour.

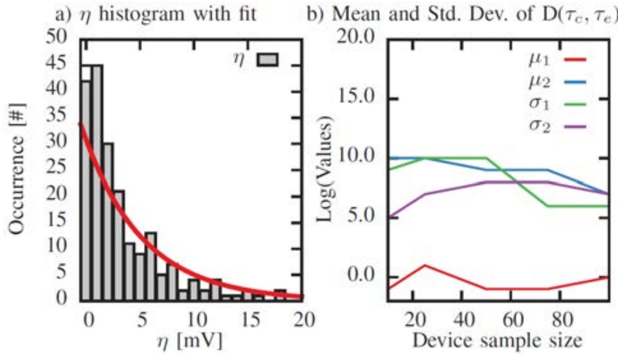


Figure 4.39: a) Histogram of η from 211 devices with 417 defects at V_{GS} = 2.5V with fit of exponential function. b) Mean and standard deviation of $D(\tau_c, \tau_e)$ as a function of device sample size.

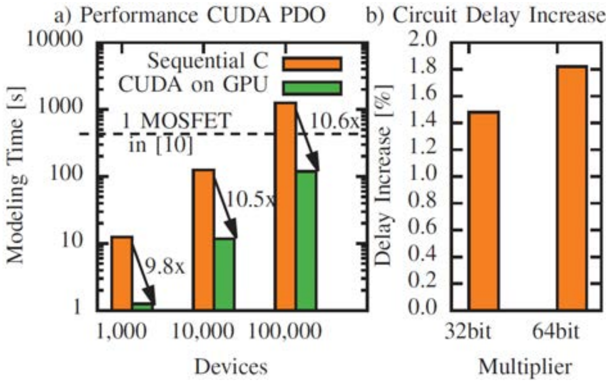


Figure 4.40: a) Performance of our CUDA implementation of our PDO model for V_{GS} waveforms with 1000 points and different number of devices. b) Circuit simulations of 32bit and 64bit multiplier synthesized in a commercial EDA tool flow resulting in 11288 respectively 42534 devices.

The presented novel parallel-BTI approach utilizing GPU-based computing exhibits a significant performance simulating, for instance, a total of 100,000 devices when compared with conventional non-parallel simulation frameworks. This methodology utilizes a calibrated version of the PDO model by means of the WTPL that analyses BTI recovery trances statistically characterized by means of the TARS software from the nanometric devices of the ENDURANCE IC chip.

4.3 TZV and TDV tests using a 28-nm IMEC IC design

As discussed in previous sections of this chapter, during real circuit operation different variability sources of deeply scaled devices associated to TZV and TDV such as BTI and HCI, are present and can potentially degrade transistors performances. In this scenario, the present section will illustrate some of the preliminary variability measurements extracted from the analysis of a new 28-nm IC designed at IMEC, during a research internship in the Device Reliability and Electrical Characterization group (DRE) group. The new IC design was utilized to capture degradation statistics of multiple device degradation metrics and account for BTI and HCI induced channel degradation. An early version of the tested 28-nm IC can be found in [127, 174].

The IC design is based on an array-based architecture containing a total of 2,500 individually controlled planar nFET and pFET transistors. A partial circuit block description is shown in Figure 4.41. The on-chip devices are distributed in a set of 250 columns where each column contains a total of 10 FETs sharing the same gate lines while each drain of each FET is connected to an individual drain biasing line. Each column in the IC allows to address electrical characterization to the 10 connected transistors simultaneously by means of individual controllable pass gates. On the contrary, the 10 IC drain lines are connected directly to the IC FETs without pass gates in between different FETs. Moreover, all the source terminals are connected together as well as the bulk connections that can be both accessed externally by two individual chip pads.

As illustrated in Figure 4.41, the on-chip devices are selected by means of a decoder utilized to activate any required number of gate lines. The decoder is controlled by a shift register and a latch intended to avoid unwanted device stress while the data is serially programmed into the shift registers. Furthermore, the circuit utilizes sense lines at the output of the shift register, and on the gate and drain of a device within the array. This allows to verify the functionality of the periphery, and to identify voltage drops on the gate and the drains. Finally, the peripheral circuits (shift registers, latches and pass gates) are entirely designed with CORE devices, as possible degradation of the pass gates will not cause substantial error on the read-out but, nevertheless, must be taken into account during the data pros-processing.

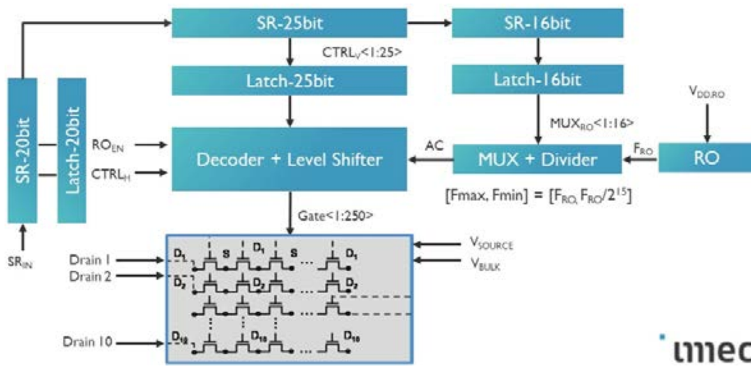


Figure 4.41: Partial circuit block definition of the tested 28-nm IC design used for BTI/HCI device testing.

Several samples of the chip have been characterized under BTI and HCI stress using the conventional MSM accelerated stress schemes. First, before the application of overvoltage stress to the on-chip devices, a complete time-zero variability characterization has been executed on all devices by means of I_{DS}–V_{GS} curve sweeps. In this sense, Figure 4.42 shows a set of seven empirical CDFs of the fresh V_{th} electrical FET parameters, two for nMOS devices and 5 for pMOS FETs where a total of 17,500 FETs have been characterized. All measurements are conducted at room temperature.

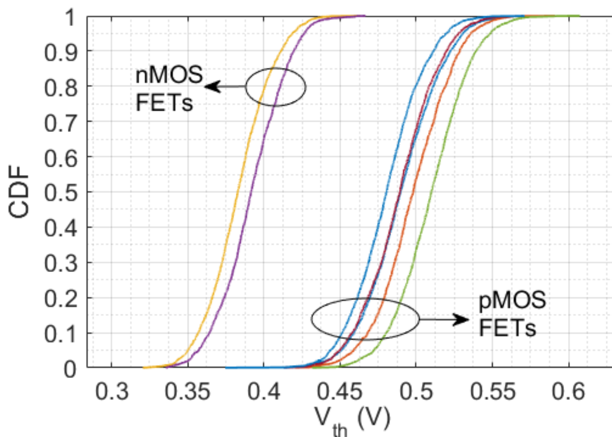


Figure 4.42: Empirical CDFs of the V_{th} electrical FETs parameters obtained from the analysis 7 IC chips.

In order to execute variability tests to the on-chip devices, a custom-built switching matrix with separately controllable voltage sources to steer the periphery and K26xx SMUs for current characterization and voltage application. After the massive extraction of the fresh parameters for each device, a sequential 5-cycle MSM scheme has been executed with stress times of 1s, 3s, 10s, 30s, and 100s, including a pre- and post-stress $I_{DS} - V_{GS}$ curve characterization period to compute device parameter shifts.

An important feature of the tested IC chip consists in a built-in ring oscillator design utilized to apply AC stress to the FETs gate terminals. The ring oscillator is accompanied by a 16x frequency divider that allows to apply different square AC frequency signals for the study of BTI and HCI damage in terms of AC stress. In this sense, Figure 4.43 (a) and (b) shows a preliminary set of chip measurement obtained after the application of HCI stress to an entire pFET chip shown in (a) and to an nFET chip shown in (b). In both cases and during all stress phases, a fixed $|V_{DS}| = 2V$ has been utilized while the $|V_{GS}|$ has been swept from 0.8V to 2V, in order to study the behaviour of the $< \Delta V_{th} >$ as a function of the increasing AC stress frequency signal for each voltage combination.

As shown in Figure 4.43 (a) and (b) a common behaviour can be observed as soon as the V_{GS} equals the fixed V_{DS} of 2V. The damage suffered by the tested devices becomes more severe with increasing V_{GS} biasing, where the $V_{DS} = V_{GS} = 2V$ combination becomes the highest harmful voltage combination in agreement with [78, 159, 160]. From the HCI stress executed to the pMOS FETs in Figure 4.43 (a) a positive $< \Delta V_{th} >$ can be observed resulting from accelerated aging damage while, for the 1.6V and 2V V_{GS} higher stress voltages, a large negative $< \Delta V_{th} >$ is shown. For the nFET HCI degradation, shown in Figure 4.43(b), a large positive $< \Delta V_{th} >$ degradation can be observed for all V_G/V_D voltage combinations.

In terms of HCI frequency behaviour, preliminary results a weak dependence can be observed in both nFET and pFET $< \Delta V_{th} >$ degradation starting approximately at 100kHz affecting the degradation trend for both pMOS and nMOS as shown in Figure 4.43(a) and (b) respectively. A possible related cause of this behaviour could be attributed to device self-heating effects during the stress as in agreement with [175, 176]. In the case of applying an AC signal below 10kHz the $< \Delta V_{th} >$ degradation appears to be closer to the DC stress probably because AC stress below 10kHz behaves thermally similar to DC stress. Nevertheless, results shown in Figure 4.43 are part of an on-going research at IMEC and more conclusive results will be reported

elsewhere.

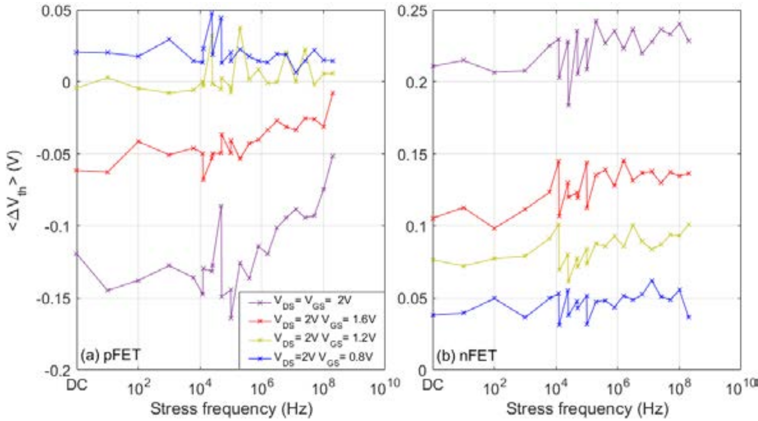


Figure 4.43: HCI versus frequency stress $< \Delta V_{th} >$ damage for pFET in (a) and for nFET in (b) is shown for four different Drain/Gate voltage pairs.

During the entire characterization executed to on-chip devices of the 28-nm IC array-based IC at IMEC facilities, BTI and HCI aging-related degradation under AC/DC stress conditions, only $I_{DS} - V_{GS}$ curves characteristics have been utilized to monitor pre- and post-stressed devices. In this sense, capturing degradation at the very first instants of the aging measurements, e.g., microsecond time scale, can be difficult since $I_{DS} - V_{GS}$ curves where utilized assuming an increasing time gap between the tested devices. In this sense, with the aim of significantly reduce the time gap between the stress and the measurement phase of an aging tests, the following section of this chapter will describe a programming solution implemented at IMEC facilities for the characterization of aging test utilizing the ultrafast unit model B1530A of the B1530 semiconductor parameters analyser.

4.4 B1530A Ultrafast device characterization

As discussed in the present chapter, when executing BTI or HCI aging tests, the elapsed time gap between the stress and the initial instants of the measurement periods becomes a critical step towards capturing the very first steps of the devices recovery after stress. In this scenario, this section

will describe the algorithm developed during a research internship at IMEC facilities and utilized to allow full control of the ultrafast B1530A Keysight module to execute stress or measurement sequences with a sampling rate as short as 10ns. Figure 4.44 shows the flux diagram of the developed C++ control software that automates the definition and execution of arbitrary variability tests and BTI/HCI accelerated MSM conventional aging sequences.

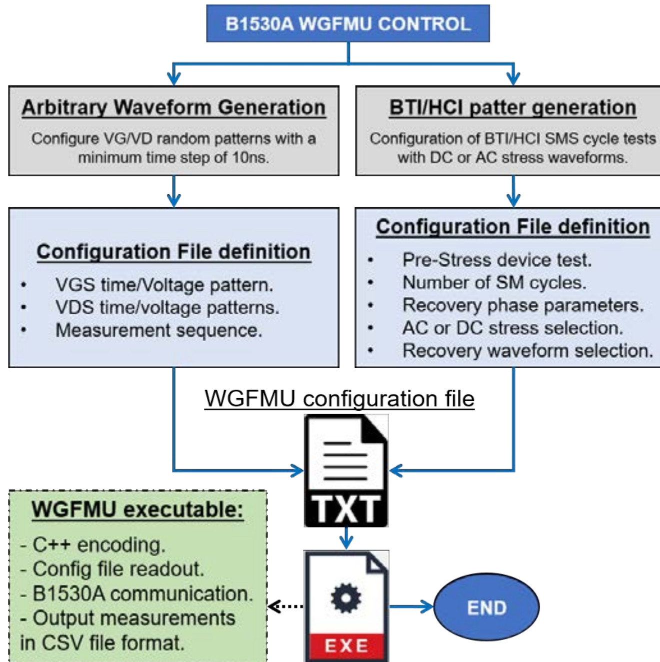


Figure 4.44: Diagram flow of the Keysight B1530A ultrafast control algorithm.

As shown in the left branch of the diagram flow of Figure 4.44, the user is able to select the arbitrary waveform option in order to execute customize variability tests. The procedure to define this kind of tests consists in entering, to the WGFMU configuration file, the V_{GS}/V_{DS} voltage/time definition patterns that will be read and executed by the WGFMU executable file. Moreover, the user can determine where to execute ultrafast measurement during the previously defined V_{GS}/V_{DS} patterns.

The second functionality of the algorithm, that corresponds to the right branch of the diagram flow in Figure 4.44 consists in the definition and exe-

cution of accelerated BTI/HCI aging MSM sequences. In order to execute an ultrafast aging test, the following definitions have to be entered in the "WGFMU configuration file": definition of the pre-stress device characterization (for instance execution of an $I_{DS} - V_{GS}$ curve), definition of the number of stress and measure or recovery cycles to be executed and the definition of the type of stress to be applied to the samples either DC or AC stress. For the case of AC stress the user can determine the type of waveform, square, triangle, etc. and also the frequency of the AC signal. Finally, the algorithm allows to determine the type of waveform to be executed during the recovery phases of the aging test, that can be defined arbitrarily.

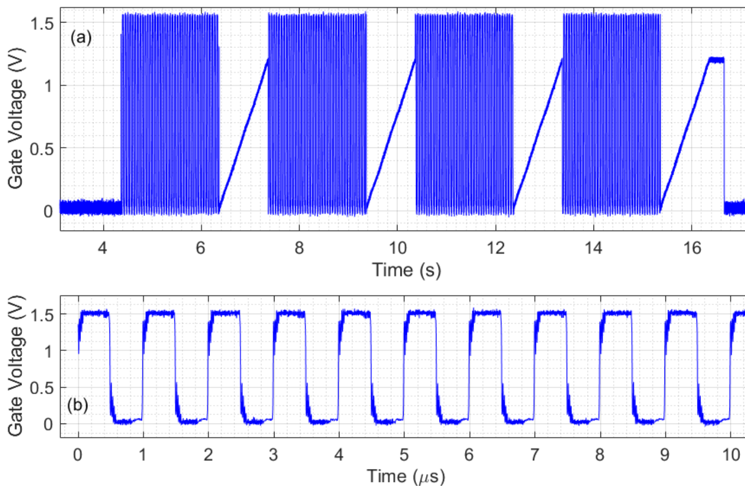


Figure 4.45: Illustration of a 4-cycle BTI aging test with 1MHz AC square signal stress followed by $I_{DS} - v_{GS}$ recovery traces in (a) 1MHz square signal

Once the "WGFMU configuration file" is fully filled with all the required variables to execute the desired tests, the "WGFMU executable file" launch the variability tests, first reading all the test variables from the configuration files, secondly, the communication with the B1500 is established by means of a GPIB connection in order to control the ultrafast module B1530A and finally, the test is executed and the data measured by the equipment is exported in a column data CSV file for its further analysis. An illustrative example of an ultrafast BTI test conducted using the described algorithm is shown in Figure 4.45, where the waveform applied to the Gate of a transistor has been captured with a digital oscilloscope.

The BTI test example shown in Figure 4.45(a) consists of a 4-Stress-Measurement (SM) cycles test where, in the four stress phases, an AC square signal oscillates between 0v and 1.5V with a frequency of 1MHz is applied to the Gate of a hypothetical device. On the other hand, during the recovery phases of the proposed BTI test a V_G sweep is executed to the Gate terminal maintaining the V_D fixed to extract the $I_{DS} - V_{GS}$ after the stress periods. Figure 4.45(b) depicts the AC square signal executed by the described algorithm verifying the square shape of the stress signal of 1MHz with a 50% duty cycle.

To illustrate the capabilities of the described ultrafast algorithm, Figure 4.46 shows the results of a 4-SM BTI test. The executed BTI test involves a single InGaAs n-type MOSFET device fabricated at IMEC facilities with a channel width of $50\mu m$ and a length of $1\mu m$. During the BTI test, the stress time periods are set to $10\mu s$, 1ms, 100ms and 1s with a DC V_{GS} of 1.5V and $V_{DS} = 0V$.

After each stress period, a I_{DS}–V_{GS} curve characterization of $1\mu s$ has been conducted to account for the degradation suffered by the device with a $V_{DS} = 50mV$ and a V_{GS} sweep from -0.5V to 0.5V. The results shown in Figure 4.46, reveal that after each accelerated stress phase, a positive ΔV_{th} appears during the recovery phase since the I_{DS}–V_{GS} curves are progressively right shifted.

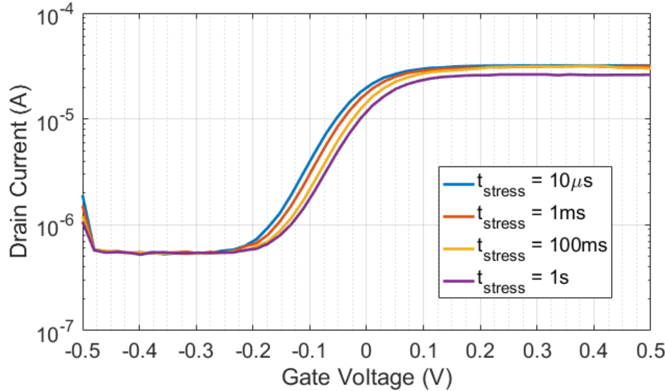


Figure 4.46: $I_{DS} - V_{GS}$ ultrafast curves characterized in a recovery time of $1\mu s$ after 4 sequential stress periods in a MSM conventional BTI test.

4.5 Conclusions

The present chapter has shown extensive variability study carried out in individual devices and a set of three novel methodologies to accurately account for the TDV related variability in nanometric MOSFET devices. In the first place, the expected lifetime prediction of the 65-nm ENDURANCE chip samples, based on statistical time-zero and time-dependent variability characterization including BTI and HCI accelerated aging tests has been described.

In the second block, a detailed description of three different methodologies utilized for the transistor's parameters extraction and reliability simulation of nanometric devices and circuits have been described. The first methodology is utilized to obtain the model card MOSFET parameters from electrical device characterization under TDV aging phenomena. The second methodology consists in an RTN-removal method utilized to get rid of any presence of RTN signals that can appear during the recovery phases of an aging test to solely account for the BTI-related behaviour. The third methodology described in this chapter, have consisted in a novel BTI-aware simulation approach that exploits the computational resources of a highly parallel, multithreaded and multicore GPU-based system coded in CUDA programming language in the sine a of research internship conducted at the KIT institute in Germany. The new framework approach utilizes the PDO model re-programmed in CUDA language to simulate thousands of transistors in large circuits like multipliers, under BTI aging conditions archiving a significant simulation time reduction.

Finally, in the last two sections of this chapter the preliminary results obtained during a research internship at IMEC facilities have been described. In the first place, TZV and TDV statistical analysis of a novel 28-nm IMEC ID design have been presented including the possibility of applying different AC stress signals to the on-chip FETs during BTI or HCI accelerated aging tests. Moreover, the main functionalities of an ultrafast characterization algorithm, developed in C++ codification, are shown utilized to apply any desired arbitrary stress waveforms together with the possibility of executing user-defined MSM sequences for BTI or HCI device testing utilizing AC/DC stress waveform with oscillation frequencies up to 1MHz.

Conclusions

The rapidly growing CMOS technology industry has become one of the major drives of the modern worldwide economy in the last decades. Since the early days of switching electronics, the key challenge of the CMOS industry has been focused in the transistors dimensions scaling in order to integrate more and more devices into the same IC chip area keeping yield and performance under control in every new technology node. Device shrink contributes to important benefits such as increasing processor speed, reducing device power consumption and reducing manufacturing cost per chip, among others. Nevertheless, the reduction of devices dimensions approaching the atomic scale size, gives rise to important variations in the transistor main parameters between identical fabricated devices at the manufacturing process, as well as, a progressive parameter variations during devices and circuit functionality over time. For this reason, the CMOS variability and reliability of nanometric devices and circuits has been intensively studied during the last decades.

In this scenario, this thesis contributes to the study of the variability and reliability of nanometric CMOS devices. The core of this thesis aims the statistical characterization of MOSFET devices to study the different sources of transistors variability. Thus, is it analyzed the time-zero variability (TZV), which is due to the device fabrication process, and the time-dependent variability, which includes transient effects due to Random Telegraph Noise (RTN) and aging mechanisms as Bias Temperature Instability (BTI) and Hot Carrier Injection (HCI).

Conventional device characterization techniques conducted using probe stations for on-wafer device measurements results in long serial aging test times when thousands of transistors are involved. In this regard, to perform fast and trustworthy statistical characterization of TZV and TDV phenomena, a novel and versatile array-based IC named ENDURANCE has been designed. The ENDURANCE IC chip contains sufficient test devices (nMOS and pMOS) including eight different transistor sizes to get statistically significant results

for any given variability test including TZV, RTN, BTI as well as HCI aging phenomenon in a unique IC design. The fabricated chip allows to execute automatic characterization of thousands of DUTs.

The IC incorporates a novel circuit design that ensures the ability to perform trustworthy device level reliability characterization. The ENDURANCE IC chip has been designed utilizing Force-&-Sense internal voltage biasing system in order to mitigate IC IR-drops and ensure a correct voltage biasing during variability testing. The IC design incorporates I/O transmission gates that allow applying external stress voltages to the DUT terminals up to 3.3V during aging tests without significant degradation of the access circuitry. Moreover, the transmission gate design also allows high DC current flow to execute HCI aging tests without incurring in electromigration damage of the chip paths. Finally, one of the main novelties of the presented ENDURANCE chip consists in the possibility of applying stress parallelization techniques which enables faster aging characterization.

With the aim of overcoming traditional on-wafer testing issues, such as serial device aging characterization resulting in extremely large test times, array-based integrated circuits have been proposed for massive device variability characterization to significantly reduce variability test times from years to hours. In this scenario, a novel and flexible characterization framework, named TARS software, has been developed, in the context of an intership in the IMSE-CNM in Sevilla (Spain), to automatically characterize the on-chip samples of array-based IC chips covering the necessary hardware and software requirements for a trustworthy and straightforward transistor characterization.

The TARS software consists in a full-custom measurement tool utilized for the automated generation and execution of variability tests to characterize the 65-nm transistors of the ENDURANCE IC that enables TZV and pipelined TDV characterization. The measurement setup permits to execute several tests: ramped voltages (I_{DS}–V_{GS}, I_{DS}–V_{DS}), RTN measurements, and BTI and HCI aging evaluations. These tests can be defined in a matter of seconds and thousands of commands are automatically generated for instrumentation and chip control. The described measurement setup in combination with the ENDURANCE IC chip, powered by the All DUT Parallel Stress Pipeline Measurements (ADPSPM) novel algorithm, allows the application of smart stress parallelization techniques to a large number of devices by overlapping the stress phases of an aging tests, reducing the measurement time from years to days.

After the execution of statistical accelerated BTI and HCI variability measurements utilizing the ENDURANCE chip together with the TARS software framework, the expected lifetime prediction as a function of the percentage of non-functional devices has been accurately obtained. Based on a maximum degradation criterion of 30mV in the threshold voltage shift, high degradation has been obtained for the NBTI case where, after the extrapolation to the nominal operation conditions, the $\approx 27\%$ of the tested devices at room temperature will became non-functional in 10 years. Moreover, for NBTI evaluated at 80 $°C$, the expected lifetime will approach to only one year. For the HCI case, the pMOS devices have shown slightly more degradation than nMOS ones and in both cases the lifetime is reduced to no more than a few months. It has been also shown that, for the 65-nm characterized technology, HCI damage is the most harmful aging degradation reducing the expected lifetime to a few days. Nevertheless, the expected lifetime obtention highly depends on the final application or circuit to be evaluated in terms of lifetime reliability. Stress conditions and degradation criteria should be appropriately adjusted for specific circuit definition.

In addition, three different transistor parameters extraction methodologies have been described. The first methodology is utilized to obtain the model card MOSFET parameters from electrical device characterization under TDV (BTI and HCI) aging phenomena for proper circuit reliability simulation utilizing commercial tools. In this sense, threshold voltage (VTH0) and mobility (U0) model card parameters have been successfully obtained from experimental measurements utilizing the mean squared error metric when comparing experimental $I_{DS} - V_{GS}$ curves with thousands of Cadence simulated $I_{DS} - V_{GS}$ curves. Moreover, a novel I_{DS}-t to $\Delta VTH0$ model card parameter variations has been developed demonstrating that BTI and RTN related fluctuations can be perfectly registered with the proposed methodology to further be included in circuit simulations.

The second methodology consists in a novel and smart RTN-removal method utilized to get rid of any presence of RTN fluctuations that can appear during the measurement phases of BTI and HCI aging tests to solely account for the aging related degradation avoiding noise masking issues. The results show that the amplitude of, for instance, the characterized BTI-related discharges amplitudes, are exponentially distributed and the effect on the threshold voltage related shift decreases with the transistor area increase. The third methodology described in this chapter, has consisted in a novel BTI-aware simulation approach that exploits the computational resources of a highly parallel, multithreaded and multicore GPU-based system coded in

CUDA programming language in the sine of a research internship conducted at the KIT institute in Germany. The new framework approach utilizes the PDO model re-programmed in CUDA language to simulate thousands of transistors in large circuits like multipliers, under BTI aging conditions archiving a significant simulation time reduction when compared with commercial EDA tools.

Finally, and thanks to a research internship conducted at IMEC facilities in Leuven, results obtained from the characterization of device TZV and TDV utilizing a novel 28-nm IMEC IC design have been presented. The IC permits the possibility of applying AC/DC stress signals to the on-chip transistors under BTI and HCI aging tests. Moreover, the main functionalities of an ultrafast characterization algorithm, developed in C++ codification, are utilized to apply any desired arbitrary stress waveforms together with the possibility of executing user-defined MSM sequences for BTI or HCI tests utilizing AC/DC stress waveform with oscillation frequencies up to 1MHz.

In conclusion, this thesis provides a deep insight into the characterization and analysis of nanometric MOSFET devices utilizing a promising array-based IC and fully automated characterization system to obtain an accurate expected lifetime prediction. Moreover, novel MOSFET parameter extraction methodologies have been presented to accurately study variability sources present in nowadays nanometric devices. Most of the results included in this thesis have been published in international journals and conferences.

Bibliography

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

[2] M. J. M. Pelgrom, H. P. Tuinhout, and M. Vertregt, "Transistor matching in analog CMOS applications," in *International Electron Devices Meeting (IEDM)*, pp. 915–918, Dec 1998.

[3] S. K. Saha, "Compact MOSFET modeling for process Variability-Aware VLSI circuit design," *IEEE Access*, vol. 2, pp. 104–115, 2014.

[4] S. K. Saha, "Modeling process variability in scaled CMOS technology," *IEEE Design Test of Computers*, vol. 27, pp. 8–16, March 2010.

[5] K. J. Kuhn, M. D. Giles, D. Becher, P. Kolar, A. Kornfeld, R. Kotlyar, S. T. Ma, A. Maheshwari, and S. Mudanai, "Process technology variation," *IEEE Transactions on Electron Devices*, vol. 58, pp. 2197–2208, Aug 2011.

[6] L. Capodieci, "From optical proximity correction to lithography-driven physical design (1996-2006) 10 years of resolution enhancement technology and the roadmap enablers for the next decade," in *Optical Microlithography XIX*, vol. 6154, p. 615401, International Society for Optics and Photonics, 2006.

[7] K. J. Kuhn, "Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS," in *IEEE International Electron Devices Meeting (IEDM)*, pp. 471–474, Dec 2007.

[8] Y. Li, S. Yu, J. Hwang, and F. Yang, "Discrete dopant fluctuations in 20-nm/15-nm-Gate planar CMOS," *IEEE Transactions on Electron Devices*, vol. 55, pp. 1449–1455, June 2008.

[9] H. Mahmoodi, S. Mukhopadhyay, and K. Roy, "Estimation of delay variations due to random-dopant fluctuations in nanoscale CMOS circuits," *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 1787–1796, Sept 2005.

[10] P. Oldiges, Q. Lin, K. Petrillo, M. Sanchez, M. Ieong, and M. Hargrove, "Modeling line edge roughness effects in sub 100 nanometer gate length devices," in *International Conference on Simulation Semiconductor Processes and Devices*, pp. 131–134, Sept 2000.

[11] S. Kaya, A. Brown, A. Asenov, D. Magot, and T. LintonI, "Analysis of statistical fluctuations due to line edge roughness in sub-0.1 μm MOSFETs," in *Simulation of Semiconductor Processes and Devices*, pp. 78–81, Springer, 2001.

[12] A. Asenov, S. Kaya, and A. R. Brown, "Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness," *IEEE Transactions on Electron Devices*, vol. 50, pp. 1254–1260, May 2003.

[13] X. Jiang, R. Wang, T. Yu, J. Chen, and R. Huang, "Investigations on line-edge roughness (LER) and line-width roughness (LWR) in nanoscale CMOS technology: Part I–Modeling and simulation method," *IEEE Transactions on Electron Devices*, vol. 60, pp. 3669–3675, Nov 2013.

[14] R. Wang, X. Jiang, T. Yu, J. Fan, J. Chen, D. Z. Pan, and R. Huang, "Investigations on line-edge roughness (LER) and line-width roughness (LWR) in nanoscale CMOS technology: Part II–Experimental results and impacts on device variability," *IEEE Transactions on Electron Devices*, vol. 60, pp. 3676–3682, Nov 2013.

[15] C. H. Diaz, H.-J. Tao, Y.-C. Ku, A. Yen, and K. Young, "An experimentally validated analytical model for gate line-edge roughness (LER) effects on technology scaling," *IEEE Electron Device Letters*, vol. 22, pp. 287–289, June 2001.

[16] A. Asenov and S. Kaya, "Effect of oxide interface roughness on the threshold voltage fluctuations in decanano MOSFETs with ultrathin gate oxides," in *International Conference on Simulation Semiconductor Processes and Devices*, pp. 135–138, Sept 2000.

[17] E. Cassan, P. Dollfus, S. Galdin, and P. Hesto, "Calculation of direct tunneling gate current through ultra-thin oxide and oxide/nitride stacks in MOSFETs and h-MOSFETs," *Microelectronics Reliability*, vol. 40, no. 4, pp. 585 – 588, 2000.

[18] R. Difrenza, J. C. Vildeuil, P. Llinares, and G. Ghibaudo, "Impact of grain number fluctuations in the MOS transistor gate on matching performance," in *International Conference on Microelectronic Test Structures*, pp. 244–249, March 2003.

[19] A. R. Brown, J. R. Watling, and A. Asenov, "Intrinsic parameter fluctuations due to random grain orientations in high-κ gate stacks," *Journal of Computational Electronics*, vol. 5, no. 4, pp. 333–336, 2006.

[20] A. Asenov and S. Saini, "Polysilicon gate enhancement of the random dopant induced threshold voltage fluctuations in sub-100 nm MOSFETs with ultrathin gate oxide," *IEEE Transactions on Electron Devices*, vol. 47, pp. 805–812, April 2000.

[21] A. R. Brown, G. Roy, and A. Asenov, "Poly-Si-Gate-Related variability in decananometer MOSFETs with conventional architecture," *IEEE Transactions on Electron Devices*, vol. 54, pp. 3056–3063, Nov 2007.

[22] S. Tan, "Challenges and performance limitations of high-k and oxynitride gate dielectrics for 90/65 nm CMOS technology," *Microelectronics Journal*, vol. 38, no. 6-7, pp. 783–786, 2007.

[23] S. Mahapatra, *Fundamentals of Bias Temperature Instability in MOS Transistors*. Springer, 2016.

[24] D. Nagy, G. Indalecio, A. J. García-Loureiro, M. A. Elmessary, K. Kalna, and N. Seoane, "FinFET versus gate-all-around nanowire FET: Performance, scaling, and variability," *IEEE Journal of the Electron Devices Society*, vol. 6, pp. 332–340, 2018.

[25] E. Ioannidis, S. Haendler, E. Josse, N. Planes, and G. Ghibaudo, "Characterization and modeling of drain current local variability in 28 and 14 nm FDSOI nMOSFETs," *Solid-State Electronics*, vol. 118, pp. 4–11, 2016.

[26] A. L. Zimpeck, Y. Aguiar, C. Meinhardt, and R. Reis, "Robustness of sub-22nm multigate devices against physical variability," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–4, May 2017.

[27] IEEE, "The international roadmap for devices and systems 2017 edition executive summary," *IRDS*, 2017.

[28] A. Rahimi, L. Benini, and R. K. Gupta, "Variability mitigation in nanometer CMOS integrated systems: A survey of techniques from

circuits to software," *Proceedings of the IEEE*, vol. 104, pp. 1410–1448, July 2016.

[29] M. Toledano-Luque, B. Kaczer, J. Franco, P. J. Roussel, M. Bina, T. Grasser, M. Cho, P. Weckx, and G. Groeseneken, "Degradation of time dependent variability due to interface state generation," in *Symposium on VLSI Technology*, pp. T190–T191, June 2013.

[30] B. Kaczer, J. Franco, P. Weckx, P. J. Roussel, V. Putcha, E. Bury, M. Simicic, A. Chasin, D. Linten, B. Parvais, *et al.*, "A brief overview of gate oxide defect properties and their relation to MOSFET instabilities and device and circuit time-dependent variability," *Microelectronics Reliability*, vol. 81, pp. 186–194, 2018.

[31] M. Toledano-Luque, B. Kaczer, P. Roussel, M. J. Cho, T. Grasser, and G. Groeseneken, "Temperature dependence of the emission and capture times of SiON individual traps after positive bias temperature stress," *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena*, vol. 29, no. 1, p. 01AA04, 2011.

[32] T. Grasser, H. Reisinger, P. . Wagner, F. Schanovsky, W. Goes, and B. Kaczer, "The time dependent defect spectroscopy (TDDS) for the characterization of the bias temperature instability," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 16–25, May 2010.

[33] T. Nagumo, K. Takeuchi, T. Hase, and Y. Hayashi, "Statistical characterization of trap position, energy, amplitude and time constants by RTN measurement of multiple individual traps," in *International Electron Devices Meeting (IEDM)*, pp. 28.3.1–28.3.4, Dec 2010.

[34] N. Tega, H. Miki, M. Yamaoka, H. Kume, T. Mine, T. Ishida, Y. Mori, R. Yamada, and K. Torii, "Impact of threshold voltage fluctuation due to random telegraph noise on scaled-down SRAM," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 541–546, April 2008.

[35] K. Takeuchi, "Impact of discrete-charge-induced variability on scaled MOS devices," *IEICE transactions on electronics*, vol. 95, no. 4, pp. 414–420, 2012.

[36] F. M. Puglisi, L. Larcher, A. Padovani, and P. Pavan, "A complete statistical investigation of RTN in HfO_2-based RRAM in high resistive state," *IEEE Transactions on Electron Devices*, vol. 62, no. 8, pp. 2606–2613, 2015.

[37] D. H. Tassis, S. G. Stavrinides, M. P. Hanias, C. G. Theodorou, G. Ghibaudo, and C. A. Dimitriadis, "Chaotic behavior of random telegraph noise in nanoscale UTBB FD-SOI MOSFETs," *IEEE Electron Device Letters*, vol. 38, pp. 517–520, April 2017.

[38] Z. Zhang, Z. Zhang, S. Guo, R. Wang, X. Wang, B. Cheng, A. Asenov, and R. Huang, "Comparative study on RTN amplitude in planar and FinFET devices," in *IEEE Electron Devices Technology and Manufacturing Conference (EDTM)*, pp. 109–110, Feb 2017.

[39] J. Zou, R. Wang, M. Luo, R. Huang, N. Xu, P. Ren, C. Liu, W. Xiong, J. Wang, J. Liu, J. Wu, W. Wong, S. Yu, H. Wu, S. Lee, and Y. Wang, "Deep understanding of AC RTN in MuGFETs through new characterization method and impacts on logic circuits," in *Symposium on VLSI Technology*, pp. T186–T187, June 2013.

[40] K. Ota, M. Saitoh, C. Tanaka, D. Matsushita, and T. Numata, "Experimental study of random telegraph noise in trigate nanowire MOSFETs," *IEEE Transactions on Electron Devices*, vol. 62, pp. 3799–3804, Nov 2015.

[41] M. Si, N. J. Conrad, S. Shin, J. Gu, J. Zhang, M. A. Alam, and P. D. Ye, "Low-frequency noise and random telegraph noise on near-ballistic III-V MOSFETs," *IEEE Transactions on Electron Devices*, vol. 62, pp. 3508–3515, Nov 2015.

[42] M. Luo, R. Wang, S. Guo, J. Wang, J. Zou, and R. Huang, "Impacts of random telegraph noise (RTN) on digital circuits," *IEEE Transactions on Electron Devices*, vol. 62, pp. 1725–1732, June 2015.

[43] K. Takeuchi, T. Nagumo, S. Yokogawa, K. Imai, and Y. Hayashi, "Single-charge-based modeling of transistor characteristics fluctuations based on statistical measurement of RTN amplitude," in *Symposium on VLSI Technology*, pp. 54–55, June 2006.

[44] R. Wang, J. Zou, X. Xu, C. Liu, J. Liu, H. Wu, Y. Wang, and R. Huang, "New observations on the AC random telegraph noise (AC RTN) in nano-MOSFETs," in *VLSI Technology, System and Application*, pp. 1–2, April 2012.

[45] T. Grasser, "Stochastic charge trapping in oxides: From random telegraph noise to bias temperature instabilities," *Microelectronics Reliability*, vol. 52, no. 1, pp. 39 – 70, 2012. 2011 Reliability of Compound Semiconductors (ROCS) Workshop.

[46] M. Waltl, W. Goes, K. Rott, H. Reisinger, and T. Grasser, "A single-trap study of PBTI in SiON nMOS transistors: Similarities and differences to the NBTI/pMOS case," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. XT.18.1–XT.18.5, June 2014.

[47] N. Ayala, J. Martin-Martinez, R. Rodriguez, M. Nafria, and X. Aymerich, "Unified characterization of RTN and BTI for circuit performance and variability simulation," in *European Solid-State Device Research Conference (ESSDERC)*, pp. 266–269, Sep. 2012.

[48] P. Saraza-Canflanca, J. Diaz-Fortuny, A. Toro-Frias, R. Castro-Lopez, E. Roca, J. Martin-Martinez, R. Rodrigucz, M. Nafria, and F. V. Fernandez, "Automated massive RTN characterization using a transistor array chip," in *15th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*, pp. 29–32, July 2018.

[49] N. A. Cintas, *Variabilitat dependent del temps per BIT i portadors calents en MOSFETs ultraescalats*. PhD thesis, Univertitat Autonoma de Barcelona, 2013.

[50] B. Kaczer, T. Grasser, P. J. Roussel, J. Franco, R. Degraeve, L. Ragnarsson, E. Simoen, G. Groeseneken, and H. Reisinger, "Origin of NBTI variability in deeply scaled pFETs," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 26–32, May 2010.

[51] S. Mahapatra and N. Parihar, "A review of NBTI mechanisms and models," *Microelectronics Reliability*, vol. 81, pp. 127 – 135, 2018.

[52] H. Reisinger, T. Grasser, W. Gustin, and C. Schlünder, "The statistical analysis of individual defects constituting NBTI and its implications for modeling DC- and AC-stress," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 7–15, May 2010.

[53] N. Kimizuka, K. Yamaguchi, K. Imai, T. Iizuka, C. T. Liu, R. C. Keller, and T. Horiuchi, "NBTI enhancement by nitrogen incorporation into ultrathin gate oxide for 0.10-/spl mu/m gate CMOS generation," in *Symposium on VLSI Technology*, pp. 92–93, June 2000.

[54] B. Kaczer, P. J. Roussel, T. Grasser, and G. Groeseneken, "Statistics of multiple trapped charges in the gate oxide of deeply scaled MOSFET Devices—Application to NBTI," *IEEE Electron Device Letters*, vol. 31, pp. 411–413, May 2010.

[55] K. T. Lee, W. Kang, E. Chung, G. Kim, H. Shim, H. Lee, H. Kim, M. Choe, N. Lee, A. Patel, J. Park, and J. Park, "Technology scaling on high-K & metal-gate FinFET BTI reliability," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 2D.1.1–2D.1.4, April 2013.

[56] N. Parihar, R. G. Southwick, U. Sharma, M. Wang, J. H. Stathis, and S. Mahapatra, "Comparison of DC and AC NBTI kinetics in RMG si and SiGe p-FinFETs," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 2D–4.1–2D–4.7, April 2017.

[57] B. Kaczer, T. Grasser, J. Martin-Martinez, E. Simoen, M. Aoulaiche, P. J. Roussel, and G. Groeseneken, "NBTI from the perspective of defect states with widely distributed time scales," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 55–60, April 2009.

[58] T. Grasser and B. Kaczer, "Negative bias temperature instability: Recoverable versus permanent degradation," in *37th European Solid-State Device Research Conference (ESSDERC)*, pp. 127–130, Sep. 2007.

[59] T. Grasser, T. Aichinger, H. Reisinger, J. Franco, P. Wagner, M. Nelhiebel, C. Ortolland, and B. Kaczer, "On the 'permanent' component of NBTI," in *IEEE International Integrated Reliability Workshop Final Report*, pp. 2–7, Oct 2010.

[60] B. Kaczer, T. Grasser, J. Roussel, J. Martin-Martinez, R. O'Connor, B. J. O'Sullivan, and G. Groeseneken, "Ubiquitous relaxation in BTI stressing—New evaluation and insights," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 20–27, April 2008.

[61] D. Varghese, D. Saha, S. Mahapatra, K. Ahmed, F. Nouri, and M. Alam, "On the dispersive versus arrhenius temperature activation of NBTI time evolution in plasma nitrided gate oxides: measurements, theory, and implications," in *IEEE International Electron Devices Meeting (IEDM)*, pp. 684–687, Dec 2005.

[62] T. Aichinger, M. Nelhiebel, S. Einspieler, and T. Grasser, "In situ poly Heater—A reliable tool for performing fast and defined temperature switches on chip," *IEEE Transactions on Device and Materials Reliability*, vol. 10, no. 1, pp. 3–8, 2010.

[63] G. Pobegen and T. Grasser, "On the distribution of NBTI time constants on a long, temperature-accelerated time scale," *IEEE Transactions on Electron Devices*, vol. 60, no. 7, pp. 2148–2155, 2013.

[64] K. Puschkarsky, H. Reisinger, C. Schlünder, W. Gustin, and T. Grasser, “Fast acquisition of activation energy maps using temperature ramps for lifetime modeling of BTI,” in *48th European Solid-State Device Research Conference (ESSDERC)*, pp. 218–221, Sep. 2018.

[65] H. Reisinger, T. Grasser, K. Ermisch, H. Nielen, W. Gustin, and C. Schlünder, “Understanding and modeling AC BTI,” in *International Reliability Physics Symposium (IRPS)*, pp. 6A.1.1–6A.1.8, April 2011.

[66] R. Fernandez, B. Kaczer, A. Nackaerts, S. Demuynck, R. Rodriguez, M. Nafria, and G. Groeseneken, “AC NBTI studied in the 1 Hz–2 GHz range on dedicated on-chip CMOS circuits,” in *International Electron Devices Meeting (IEDM)*, pp. 1–4, Dec 2006.

[67] T. Grasser, B. Kaczer, H. Reisinger, P. . Wagner, and M. Toledano-Luque, “On the frequency dependence of the bias temperature instability,” in *IEEE International Reliability Physics Symposium (IRPS)*, pp. XT.8.1–XT.8.7, April 2012.

[68] G. Wirth, J. Franco, and B. Kaczer, “A unified model for AC bias temperature instability,” in *IEEE International Integrated Reliability Workshop Final Report*, pp. 84–87, Oct 2013.

[69] J. Franco, B. Kaczer, M. Toledano-Luque, P. J. Roussel, J. Mitard, L. Ragnarsson, L. Witters, T. Chiarella, M. Togo, N. Horiguchi, G. Groeseneken, M. F. Bukhori, T. Grasser, and A. Asenov, “Impact of single charged gate oxide defects on the performance and scaling of nanoscaled FETs,” in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 5A.4.1–5A.4.6, April 2012.

[70] R. Gao, Z. Ji, A. B. Manut, J. F. Zhang, J. Franco, S. W. M. Hatta, W. D. Zhang, B. Kaczer, D. Linten, and G. Groeseneken, “NBTI-Generated defects in nanoscaled devices: Fast characterization methodology and modeling,” *IEEE Transactions on Electron Devices*, vol. 64, pp. 4011–4017, Oct 2017.

[71] B. Kaczer, J. Franco, P. Weckx, P. J. Roussel, V. Putcha, E. Bury, M. Simicic, A. Chasin, D. Linten, B. Parvais, *et al.*, “A brief overview of gate oxide defect properties and their relation to MOSFET instabilities and device and circuit time-dependent variability,” *Microelectronics Reliability*, vol. 81, pp. 186–194, 2018.

[72] E. Amat, T. Kauerauf, R. Degraeve, R. Rodriguez, M. Nafria, X. Aymerich, and G. Groeseneken, "Competing degradation mechanisms in short-channel transistors under channel hot-carrier stress at elevated temperatures," *IEEE Transactions on Device and Materials Reliability*, vol. 9, pp. 454–458, Sept 2009.

[73] N. D. Arora, *MOSFET models for VLSI circuit simulation: theory and practice.* Springer Science & Business Media, 2012.

[74] M. Cho, P. Roussel, B. Kaczer, R. Degraeve, J. Franco, M. Aoulaiche, T. Chiarella, T. Kauerauf, N. Horiguchi, and G. Groeseneken, "Channel hot carrier degradation mechanism in Long/Short channel n-FinFETs," *IEEE Transactions on Electron Devices*, vol. 60, pp. 4002–4007, Dec 2013.

[75] A. Bayer, M. Porti, J. Martin-Martínez, M. Lanza, R. Rodriguez, V. Velayudhan, E. Amat, M. Nafria, X. Aymerich, M. B. Gonzalez, and E. Simoen, "Channel hot-carriers degradation in MOSFETs: A conductive AFM study at the nanoscale," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 5D.4.1–5D.4.6, April 2013.

[76] G. Groeseneken, R. Degraeve, B. Kaczer, and K. Martens, "Trends and perspectives for electrical characterization and reliability assessment in advanced CMOS technologies," in *Proceedings of the European Solid State Device Research Conference*, pp. 64–72, Sept 2010.

[77] A. Bravaix, V. Huard, D. Goguenheim, and E. Vincent, "Hot-carrier to cold-carrier device lifetime modeling with temperature for low power 40nm si-bulk NMOS and PMOS FETs," in *International Electron Devices Meeting (IEDM)*, pp. 27.5.1–27.5.4, Dec 2011.

[78] M. Duan, J. F. Zhang, Z. Ji, W. D. Zhang, B. Kaczer, and A. Asenov, "Key issues and solutions for characterizing hot carrier aging of nanometer scale nMOSFETs," *IEEE Transactions on Electron Devices*, vol. 64, pp. 2478–2484, June 2017.

[79] J. Fang and S. S. Sapatnekar, "Incorporating hot-carrier injection effects into timing analysis for large circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 12, pp. 2738–2751, 2014.

[80] U. Sharma and S. Mahapatra, "A SPICE compatible compact model for hot-carrier degradation in MOSFETs under different experimental conditions," *IEEE Transactions on Electron Devices*, vol. 66, pp. 839–846, Feb 2019.

[81] S. Tam, P.-K. Ko, and C. Hu, “Lucky-electron model of channel hot-electron injection in MOSFET’s,” *IEEE Transactions on Electron Devices*, vol. 31, no. 9, pp. 1116–1125, 1984.

[82] S. E. Rauch and G. La Rosa, “The energy-driven paradigm of NMOSFET hot-carrier effects,” *IEEE Transactions on Device and Materials Reliability*, vol. 5, no. 4, pp. 701–705, 2005.

[83] Y. M. Randriamihaja, A. Zaka, V. Huard, M. Rafik, D. Rideau, D. Roy, A. Bravaix, and P. Palestri, “Hot carrier degradation: From defect creation modeling to their impact on NMOS parameters,” in *IEEE International Reliability Physics Symposium (IRPS)*, pp. XT.15.1–XT.15.4, April 2012.

[84] C. Guerin, V. Huard, and A. Bravaix, “The energy-driven hot-carrier degradation modes of nMOSFETs,” *IEEE Transactions on Device and Materials Reliability*, vol. 7, no. 2, pp. 225–235, 2007.

[85] A. Spessot, M. Aoulaiche, M. Cho, J. Franco, T. Schram, R. Ritzenthaler, and B. Kaczer, “Impact of off state stress on advanced high-K metal gate NMOSFETs,” in *44th European Solid State Device Research Conference (ESSDERC)*, pp. 365–368, Sep. 2014.

[86] M. Cho, A. Spessot, B. Kaczer, M. Aoulaiche, R. Ritzenthaler, T. Schram, P. Fazan, N. Horiguchi, and D. Linten, “Off-state stress degradation mechanism on advanced p-MOSFETs,” in *International Conference on IC Design Technology (ICICDT)*, pp. 1–4, June 2015.

[87] K. O. Jeppson and C. M. Svensson, “Negative bias stress of MOS devices at high electric fields and degradation of MNOS devices,” *Journal of Applied Physics*, vol. 48, no. 5, pp. 2004–2014, 1977.

[88] H. Kükner, S. Khan, P. Weckx, P. Raghavan, S. Hamdioui, B. Kaczer, F. Catthoor, L. V. der Perre, R. Lauwereins, and G. Groeseneken, “Comparison of Reaction-Diffusion and atomistic Trap-Based BTI models for logic gates,” *IEEE Transactions on Device and Materials Reliability*, vol. 14, pp. 182–193, March 2014.

[89] S. Ogawa and N. Shiono, “Generalized diffusion-reaction model for the low-field charge-buildup instability at the Si-SiO_2 interface,” *Phys. Rev. B*, vol. 51, pp. 4218–4230, Feb 1995.

[90] B. Kaczer, V. Arkhipov, R. Degraeve, N. Collaert, G. Groeseneken, and M. Goodwin, “Disorder-controlled-kinetics model for negative bias temperature instability and its experimental verification,” in *IEEE*

International Reliability Physics Symposium (IRPS), pp. 381–387, April 2005.

[91] T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P. Wagner, F. Schanovsky, J. Franco, M. T. Luque, and M. Nelhiebel, "The paradigm shift in understanding the bias temperature instability: From Reaction–Diffusion to switching oxide traps," *IEEE Transactions on Electron Devices*, vol. 58, pp. 3652–3666, Nov 2011.

[92] T. Grasser, B. Kaczer, W. Goes, T. Aichinger, P. Hehenberger, and M. Nelhiebel, "A two-stage model for negative bias temperature instability," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 33–44, April 2009.

[93] S. Gupta, B. Jose, K. Joshi, A. Jain, M. A. Alam, and S. Mahapatra, "A comprehensive and critical re-assessment of 2-stage energy level NBTI model," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. XT.3.1–XT.3.6, April 2012.

[94] V. Huard, "Two independent components modeling for negative bias temperature instability," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 33–42, May 2010.

[95] G. Rzepa, J. Franco, A. Subirats, M. Jech, A. Chasin, A. Grill, M. Waltl, T. Knobloch, B. Stampfer, T. Chiarella, N. Horiguchi, L. . Ragnarsson, D. Linten, B. Kaczer, and T. Grasser, "Efficient physical defect model applied to PBTI in high-k stacks," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. XT–11.1–XT–11.6, April 2017.

[96] T. Grasser, W. Goes, Y. Wimmer, F. Schanovsky, G. Rzepa, M. Waltl, K. Rott, H. Reisinger, V. V. Afanas'ev, A. Stesmans, A. El-Sayed, and A. L. Shluger, "On the microscopic structure of hole traps in pMOSFETs," in *IEEE International Electron Devices Meeting (IEDM)*, pp. 21.1.1–21.1.4, Dec 2014.

[97] T. Grasser, P. Wagner, H. Reisinger, T. Aichinger, G. Pobegen, M. Nelhiebel, and B. Kaczer, "Analytic modeling of the bias temperature instability using capture/emission time maps," in *International Electron Devices Meeting (IEDM)*, pp. 27.4.1–27.4.4, Dec 2011.

[98] J. Martin-Martinez, B. Kaczer, M. Toledano-Luque, R. Rodriguez, M. Nafria, X. Aymerich, and G. Groeseneken, "Probabilistic defect occupancy model for NBTI," in *International Reliability Physics Symposium (IRPS)*, pp. XT.4.1–XT.4.6, April 2011.

[99] S. Mahapatra, N. Goel, S. Desai, S. Gupta, B. Jose, S. Mukhopadhyay, K. Joshi, A. Jain, A. E. Islam, and M. A. Alam, "A comparative study of different Physics-Based NBTI models," *IEEE Transactions on Electron Devices*, vol. 60, pp. 901–916, March 2013.

[100] M. Ershov, S. Saxena, H. Karbasi, S. Winters, S. Minehane, J. Babcock, R. Lindley, P. Clifton, M. Redford, and A. Shibkov, "Dynamic recovery of negative bias temperature instability in p-type metal–oxide–semiconductor field-effect transistors," *Applied physics letters*, vol. 83, no. 8, pp. 1647–1649, 2003.

[101] S. Rangan, N. Mielke, and E. C. C. Yeh, "Universal recovery behavior of negative bias temperature instability PMOSFETs," in *IEEE International Electron Devices Meeting (IEDM)*, pp. 14.3.1–14.3.4, Dec 2003.

[102] M. Moras, J. Martin-Martinez, V. Velayudhan, R. Rodriguez, M. Nafria, X. Aymerich, and E. Simoen, "Negative bias temperature instabilities in pMOSFETs: Ultrafast characterization and modelling," in *10th Spanish Conference on Electron Devices (CDE)*, pp. 1–4, Feb 2015.

[103] E. N. Kumar, V. D. Maheta, S. Purawat, A. E. Islam, C. Olsen, K. Ahmed, M. A. Alam, and S. Mahapatra, "Material dependence of NBTI physical mechanism in silicon oxynitride (SiON) p-MOSFETs: A comprehensive study by Ultra-Fast On-The-Fly (uf-otf) IDLIN technique," in *IEEE International Electron Devices Meeting (IEDM)*, pp. 809–812, Dec 2007.

[104] M. Denais, C. Parthasarathy, G. Ribes, Y. Rey-Tauriac, N. Revil, A. Bravaix, V. Huard, and F. Perrier, "On-the-fly characterization of NBTI in ultra-thin gate oxide PMOSFET's," in *IEEE International Electron Devices Meeting (IEDM)*, pp. 109–112, Dec 2004.

[105] T. Grasser, P. Wagner, P. Hehenberger, W. Goes, and B. Kaczer, "A rigorous study of measurement techniques for negative bias temperature instability," *IEEE Transactions on Device and Materials Reliability*, vol. 8, pp. 526–535, Sept 2008.

[106] T. Sato, T. Kozaki, T. Uezono, H. Tsutsui, and H. Ochi, "A device array for efficient bias-temperature instability measurements," in *European Solid-State Device Research Conference (ESSDERC)*, pp. 143–146, Sept 2011.

[107] T. Grasser, H. Reisinger, W. Goes, T. Aichinger, P. Hehenberger, P. . Wagner, M. Nelhiebel, J. Franco, and B. Kaczer, "Switching oxide

traps as the missing link between negative bias temperature instability and random telegraph noise," in *IEEE International Electron Devices Meeting (IEDM)*, pp. 1–4, Dec 2009.

[108] A. J. Lelis and T. R. Oldham, "Time dependence of switching oxide traps," *IEEE Transactions on Nuclear Science*, vol. 41, pp. 1835–1843, Dec 1994.

[109] F. M. Puglisi, P. Pavan, A. Padovani, L. Larcher, and G. Bersuker, "RTS noise characterization of HfOx RRAM in high resistive state," *Solid-State Electronics*, vol. 84, pp. 160–166, 2013.

[110] T. Nagumo, K. Takeuchi, S. Yokogawa, K. Imai, and Y. Hayashi, "New analysis methods for comprehensive understanding of random telegraph noise," in *IEEE International Electron Devices Meeting (IEDM)*, pp. 1–4, Dec 2009.

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

[112] J. Martin-Martinez, J. Diaz, R. Rodriguez, M. Nafria, and X. Aymerich, "New weighted time lag method for the analysis of random telegraph signals," *IEEE Electron Device Letters*, vol. 35, pp. 479–481, April 2014.

[113] O. Michael, M. Juettner, and J. Hoentschel, "Mass data analysis of random telegraph noise in 22nm FDSOI back biased transistors," *Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, vol. 34, pp. 1–4, 2017.

[114] J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez, R. Castro-Lopez, E. Roca, X. Aragones, E. Barajas, D. Mateo, F. V. Fernandez, and M. Nafria, "A versatile CMOS transistor array IC for the statistical characterization of time-zero variability, RTN, BTI, and HCI," *IEEE Journal of Solid-State Circuits*, pp. 1–13, 2018.

[115] V. Putcha, M. Simicic, P. Weckx, B. Parvais, J. Franco, B. Kaczer, D. Linten, D. Verkest, A. Thean, and G. Groeseneken, "Smart-array for pipelined BTI characterization," in *IEEE International Integrated Reliability Workshop (IIRW)*, pp. 95–98, Oct 2015.

[116] V. Putcha, E. Bury, P. Weckx, J. Franco, B. Kaczer, and G. Groeseneken, "Design and simulation of on-chip circuits for parallel characterization

of ultrascaled transistors for BTI reliability," in *IEEE International Integrated Reliability Workshop Final Report (IIRW)*, pp. 99–102, Oct 2014.

[117] H. Awano, M. Hiromoto, and T. Sato, "BTIarray: A time-overlapping transistor array for efficient statistical characterization of bias temperature instability," *IEEE Transactions on Device and Materials Reliability*, vol. 14, pp. 833–843, Sept 2014.

[118] Keysight Technoloiges, "The Parametric Measurement HandBook 4th edition," 2018.

[119] C. S. Chen, L. Li, Q. Lim, H. H. Teh, N. F. B. Omar, C. Ler, and J. T. Watt, "A compact test structure for characterizing transistor variability beyond 3σ," *IEEE Transactions on Semiconductor Manufacturing*, vol. 28, pp. 329–336, Aug 2015.

[120] P. Weckx, B. Kaczer, C. Chen, J. Franco, E. Bury, K. Chanda, J. Watt, P. J. Roussel, F. Catthoor, and G. Groeseneken, "Characterization of time-dependent variability using 32k transistor arrays in an advanced HK/MG technology," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 3B.1.1–3B.1.6, April 2015.

[121] P. Weckx, B. Kaczer, C. Chen, P. Raghavan, D. Linten, and A. Mocuta, "Relaxation of time-dependent NBTI variability and separation from RTN," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. XT–9.1–XT–9.5, April 2017.

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

[123] A. Whitcombe, S. Taylor, M. Denham, V. Milovanović, and B. Nikolić, "On-chip I–V variability and random telegraph noise characterization in 28 nm CMOS," in *46th European Solid-State Device Research Conference (ESSDERC)*, pp. 248–251, Sept 2016.

[124] M. Simicic, V. Putcha, B. Parvais, P. Weckx, B. Kaczer, G. Groeseneken, G. Gielen, D. Linten, and A. Thean, "Advanced MOSFET variability and reliability characterization array," in *IEEE International Integrated Reliability Workshop (IIRW)*, pp. 73–76, Oct 2015.

[125] C. Schlünder, J. M. Berthold, M. Hoffmann, J. . Weigmann, W. Gustin, and H. Reisinger, "A new smart device array structure for statistical investigations of BTI degradation and recovery," in *International Reliability Physics Symposium (IRPS)*, pp. 2B.6.1–2B.6.5, April 2011.

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

[127] E. Bury, B. Kaczer, K. Chuang, J. Franco, P. Weckx, A. Chasin, M. Simicic, D. Linten, and G. Groeseneken, "Statistical assessment of the full VG/VD degradation space using dedicated device arrays," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 2D–5.1–2D–5.6, April 2017.

[128] M. B. da Silva, B. Kaczer, G. V. der Plas, G. I. Wirth, and G. Groeseneken, "On-chip circuit for massively parallel BTI characterization," in *IEEE International Integrated Reliability Workshop Final Report*, pp. 90–93, Oct 2011.

[129] J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez, M. Nafria, R. Castro-Lopez, E. Roca, F. V. Fernandez, E. Barajas, X. Aragones, and D. Mateo, "A transistor array chip for the statistical characterization of process variability, RTN and BTI/CHC aging," in *14th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*, pp. 1–4, June 2017.

[130] 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*, pp. 1–13, 2019.

[131] J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez, M. Nafria, R. Castro-Lopez, E. Roca, and F. V. Fernandez, "TARS: A toolbox for statistical reliability modeling of CMOS devices," in *14th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*, pp. 1–4, June 2017.

[132] S. Gerardin, M. Bagatin, D. Cornale, L. Ding, S. Mattiazzo, A. Paccagnella, F. Faccio, and S. Michelis, "Enhancement of transistor-to-transistor variability due to total dose effects in 65-nm MOSFETs," *IEEE Transactions on Nuclear Science*, vol. 62, no. 6, pp. 2398–2403, 2015.

[133] G. Chen, T. Fujimura, Q. Dong, S. Nakatake, and B. Yang, "DC characteristics and variability on 90nm CMOS transistor array-style analog layout," *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 21, no. 3, p. 45, 2016.

[134] A. S. Sedra and K. C. Smith, *Microelectronic Circuits*. Oxford University Press, 2014.

[135] R. J. Baker, *CMOS: circuit design, layout, and simulation*, vol. 1. John Wiley & Sons, 2010.

[136] C. U. N. S. SLEEP, "Minimizing the sub threshold leakage for high performance CMOS circuits using novel stacked sleep transistor technique," *International Journal of Electrical and Electronic Engineering and Telecomunications*, 2012.

[137] M. Sangeetha, S. Selvarani, N. Indhumathi, and M. Ramya., "Design of low power modified wilson current mirror based level shifter using forced PMOS method," *International Journal of Modern Trends in Engineering and Research*, 2015.

[138] T. Grasser, W. Gos, V. Sverdlov, and B. Kaczer, "The universality of NBTI relaxation and its implications for modeling and characterization," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 268–280, April 2007.

[139] IEEE, "IEEE standard codes; formats; protocols; and common commands. for use with ANSI/IEEE std 488.1-1987 IEEE standard digital interface for programmable instrumentation," *ANSI/IEEE Std 488.2-1987*, pp. 0–1, 1988.

[140] National Instruments, "NI USB-6501 User's Guide," 2015.

[141] Keysight Technoloiges, "Keysight E3631A Triple Output DC Power Supply User's Guide," 2013.

[142] Keysight Technoloiges, "Keysight Technologies B1500A/B1505A/B1506A/B1507A Device Analyzer Series Programming Guide," 2016.

[143] Thermonics, "T-2650BV Precision Temperature Forcing System User's and Maintenance Manual," 2007.

[144] S. H. Voldman, *ESD: circuits and devices*. John Wiley & Sons, 2015.

[145] ANALOG DEVICES, “3.0 kV RMS/3.75 kV RMS Quad Digital Isolators Data Sheet,” 2018.

[146] D. K. Schroder, *Semiconductor material and device characterization.* John Wiley & Sons, 2006.

[147] W. Liu and C. Hu, *BSIM4 and MOSFET Modeling for IC simulation.* World Scientific, 2011.

[148] A. Ortiz-Conde, A. Sucre-Gonzalez, F. Zarate-Rincon, R. Torres-Torres, R. S. Murphy-Arteaga, J. J. Liou, and F. J. García-Sánchez, “A review of DC extraction methods for MOSFET series resistance and mobility degradation model parameters,” *Microelectronics Reliability*, vol. 69, pp. 1–16, 2017.

[149] A. Ortiz-Conde, F. J. Garcia-Sanchez, J. Muci, A. T. Barrios, J. J. Liou, and C.-S. Ho, “Revisiting MOSFET threshold voltage extraction methods,” *Microelectronics Reliability*, vol. 53, no. 1, pp. 90 – 104, 2013. Reliability of Micro-Interconnects in 3D IC Packages.

[150] M. Tsuno, M. Suga, M. Tanaka, K. Shibahara, M. Miura-Mattausch, and M. Hirose, “Physically-based threshold voltage determination for MOSFET’s of all gate lengths,” *IEEE Transactions on Electron Devices*, vol. 46, pp. 1429–1434, July 1999.

[151] P. G. Drennan and C. C. McAndrew, “Understanding MOSFET mismatch for analog design,” *IEEE Journal of Solid-State Circuits*, vol. 38, pp. 450–456, March 2003.

[152] C. Couso, J. Diaz-Fortuny, J. Martin-Martinez, M. Porti, R. Rodriguez, M. Nafria, F. V. Fernandez, E. Roca, R. Castro-Lopez, E. Barajas, D. Mateo, and X. Aragones, “Dependence of MOSFETs threshold voltage variability on channel dimensions,” in *Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, pp. 87–90, April 2017.

[153] K. Sakakibara, Y. Miura, T. Kumamoto, and S. Tanimoto, “Analysis of deviation from pelgrom scaling law in Vth variability of pocket-implanted MOSFET,” in *Custom Integrated Circuits Conference*, pp. 1–4, Sep. 2013.

[154] T. B. Hook, J. B. Johnson, A. Cathignol, A. Cros, and G. Ghibaudo, “Comment on “channel length and threshold voltage dependence of a transistor mismatch in a 32-nm HKMG technology”,” *IEEE Transactions on Electron Devices*, vol. 58, pp. 1255–1256, April 2011.

[155] M. Maestro, J. Diaz, A. Crespo-Yepes, M. Gonzalez, J. Martin-Martinez, R. Rodriguez, M. Nafria, F. Campabadal, and X. Aymerich, "New high resolution random telegraph noise (RTN) characterization method for resistive RAM," *Solid-State Electronics*, vol. 115, pp. 140–145, 2016.

[156] J. Diaz-Fortuny, M. Maestro, J. Martin-Martinez, A. Crespo-Yepes, R. Rodriguez, M. Nafria, and X. Aymerich, "Current-limiting and ultrafast system for the characterization of resistive random access memories," *Review of Scientific Instruments*, vol. 87, no. 6, p. 064705, 2016.

[157] P. Saraza-Canflanca, J. Diaz-Fortuny, R. Castro-Lopez, E. Roca, J. Martin-Martinez, R. Rodriguez, M. Nafria, and 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)*, pp. 29–32, March 2019.

[158] T. Grasser, K. Rott, H. Reisinger, M. Waltl, J. Franco, and B. Kaczer, "A unified perspective of RTN and BTI," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 4A.5.1–4A.5.7, June 2014.

[159] S. Tyaginov, M. Bina, J. Franco, D. Osintsev, O. Triebl, B. Kaczer, and T. Grasser, "Physical modeling of hot-carrier degradation for short- and long-channel MOSFETs," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. XT.16.1–XT.16.8, June 2014.

[160] M. Bina, S. Tyaginov, J. Franco, K. Rupp, Y. Wimmer, D. Osintsev, B. Kaczer, and T. Grasser, "Predictive hot-carrier modeling modeling of n-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. 61, no. 9, pp. 3103–3110, 2014.

[161] Y. Hu, D. Ang, and Z. Teo, "Threshold voltage and mobility extraction by ultrafast switching measurement on NBTI," *IEEE Transactions on Electron Devices*, vol. 57, no. 8, pp. 2027–2031, 2010.

[162] T. Rudenko, V. Kilchytska, M. K. M. Arshad, J. Raskin, A. Nazarov, and D. Flandre, "On the MOSFET threshold voltage extraction by transconductance and transconductance-to-current ratio change methods: Part I—Effect of gate-voltage-dependent mobility," *IEEE Transactions on Electron Devices*, vol. 58, pp. 4172–4179, Dec 2011.

[163] T. Rudenko, V. Kilchytska, M. K. Md Arshad, J. Raskin, A. Nazarov, and D. Flandre, "On the MOSFET threshold voltage extraction by

transconductance and transconductance-to-current ratio change methods: Part II—Effect of drain voltage," *IEEE Transactions on Electron Devices*, vol. 58, pp. 4180–4188, Dec 2011.

[164] J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez, M. Nafria, R. Castro-Lopez, E. Roca, and F. V. Fernandez, "A noise and RTN-removal smart method for parameters extraction of CMOS aging compact models," in *Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, pp. 1–4, March 2018.

[165] J. Diaz-Fortuny, J. Martin-Martinez, R. Rodriguez, R. Castro-Lopez, E. Roca, F. V. Fernandez, and M. Nafria, "A smart noise- and RTN-removal method for parameter extraction of CMOS aging compact models," *Solid-State Electronics*, 2019.

[166] 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 *15th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*, pp. 53–56, July 2018.

[167] V. M. van Santen, J. Diaz-Fortuny, H. Amrouch, J. Martin-Martinez, R. Rodriguez, R. Castro-Lopez, E. Roca, F. V. Fernandez, J. Henkel, and M. Nafria, "Weighted time lag plot defect parameter extraction and GPU-based BTI modeling for BTI variability," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. P–CR.6–1–P–CR.6–6, March 2018.

[168] P. Martín-Lloret, A. Toro-Frías, R. Castro-López, E. Roca, F. V. Fernández, J. Martin-Martinez, R. Rodriguez, and M. Nafria, "CASE: A reliability simulation tool for analog ICs," in *14th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*, pp. 1–4, June 2017.

[169] N. Ayala, J. Martin-Martinez, E. Amat, M. Gonzalez, P. Verheyen, R. Rodriguez, M. Nafria, X. Aymerich, and E. Simoen, "NBTI related time-dependent variability of mobility and threshold voltage in pMOSFETs and their impact on circuit performance," *Microelectronic Engineering*, vol. 88, no. 7, pp. 1384–1387, 2011.

[170] A. Toro-Frías, P. Martín-Lloret, J. Martin-Martinez, R. Castro-López, E. Roca, R. Rodriguez, M. Nafria, and F. Fernández, "Reliability simulation for analog ics: Goals, solutions, and challenges," *Integration*, vol. 55, pp. 341 – 348, 2016.

[171] B. Tudor, J. Wang, C. Sun, Z. Chen, Z. Liao, R. Tan, W. Liu, and F. Lee, "MOSRA: An efficient and versatile MOS aging modeling and reliability analysis solution for 45nm and below," in *10th IEEE International Conference on Solid-State and Integrated Circuit Technology*, pp. 1645–1647, Nov 2010.

[172] N. Goel, T. Naphade, and S. Mahapatra, "Combined trap generation and transient trap occupancy model for time evolution of NBTI during DC multi-cycle and AC stress," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 4A.3.1–4A.3.7, April 2015.

[173] V. M. van Santen, H. Amrouch, J. Martin-Martinez, M. Nafria, and J. Henkel, "Designing guardbands for instantaneous aging effects," in *53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 1–6, June 2016.

[174] E. Bury, A. Chasin, B. Kaczer, K. . Chuang, J. Franco, M. Simicic, P. Weckx, and D. Linten, "Self-heating-aware CMOS reliability characterization using degradation maps," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 2A.3–1–2A.3–6, March 2018.

[175] S. Mittl and F. Guarín, "Self-heating and its implications on hot carrier reliability evaluations," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 4A.4.1–4A.4.6, April 2015.

[176] C. Prasad, S. Ramey, and L. Jiang, "Self-heating in advanced CMOS technologies," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 6A–4.1–6A–4.7, April 2017.