Analysis of ultrathin gate-oxide breakdown mechanisms and applications to antifuse memories fabricated in advanced CMOS processes

Matthieu Deloge

N° d'ordre 2011-ISAL-0097 Année 2011

THÈSE

préparée à **STMicroelectronics** dans l'équipe Fuse Solutions, Crolles, France
et au **Laboratoire Ampere**, INSA Lyon

Analysis of ultrathin gate-oxide breakdown mechanisms and applications to antifuse memories fabricated in advanced CMOS processes

Contribution à l'analyse des mécanismes de claquage d'oxyde ultra mince et applications aux mémoires antifusibles en technologies avancées

présentée devant
L'Institut National des Sciences Appliquées de Lyon
pour obtenir le grade de docteur
Ecole Doctorale: **Electronique, Electrotechnique, Automatique**
Spécialité : **Energie et systèmes**
par

Matthieu Deloge

Soutenue le 15/12/2011 devant la commission d'examen

Jury

Jean Michel Portal	PR, IM2NP, Marseille	Président
Paolo Pavan	PR, Université de Modene, Italie	Rapporteur
Gérard Ghibaudo	PR, IMEP-LAHC, Grenoble	Rapporteur
Bruno Allard	PR, Laboratoire Ampère INSA, Lyon	Directeur de thèse
Philippe Candelier	ING STMicroelectronics, Crolles	Examinateur
Joël Damiens	ING STMicroelectronics, Crolles	Examinateur

INSA Direction de la Recherche - Ecoles Doctorales – Quinquennal 2011-2015

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **http://www.edchimie-lyon.fr** Insa : R. GOURDON	M. Jean Marc LANCELIN Université de Lyon – Collège Doctoral Bât ESCPE 43 bd du 11 novembre 1918 69622 VILLEURBANNE Cedex Tél : 04.72.43 13 95 directeur@edchimie-lyon.fr
E.E.A.	**ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE** **http://edeea.ec-lyon.fr** Secrétariat : M.C. HAVGOUDOUKIAN eea@ec-lyon.fr	M. Gérard SCORLETTI Ecole Centrale de Lyon 36 avenue Guy de Collongue 69134 ECULLY Tél : 04.72.18 60 97 Fax : 04 78 43 37 17 Gerard.scorletti@ec-lyon.fr
E2M2	**EVOLUTION, ECOSYSTEME, MICROBIOLOGIE, MODELISATION** **http://e2m2.universite-lyon.fr** Insa : H. CHARLES	Mme Gudrun BORNETTE CNRS UMR 5023 LEHNA Université Claude Bernard Lyon 1 Bât Forel 43 bd du 11 novembre 1918 69622 VILLEURBANNE Cédex Tél : 04.72.43.12.94 e2m2@biomserv.univ-lyon1.fr
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTE** **http://ww2.ibcp.fr/ediss** Sec : Safia AIT CHALAL Insa : M. LAGARDE	M. Didier REVEL Hôpital Louis Pradel Bâtiment Central 28 Avenue Doyen Lépine 69677 BRON Tél : 04.72.68 49 09 Fax :04 72 35 49 16 Didier.revel@creatis.uni-lyon1.fr
INFOMATHS	**INFORMATIQUE ET MATHEMATIQUES** http://infomaths.univ-lyon1.fr	M. Johannes KELLENDONK Université Claude Bernard Lyon 1 LIRIS - INFOMATHS Bâtiment Nautibus 43 bd du 11 novembre 1918 69622 VILLEURBANNE Cedex Tél : 04.72. 43.19.05 Fax 04 72 43 13 10 infomaths@bat710.univ-lyon1.fr
Matériaux	**MATERIAUX DE LYON**	M. Jean-Yves BUFFIERE Secrétaire : Mériem LABOUNE INSA de Lyon École Doctorale Matériaux Mériem LABOUNE Bâtiment Antoine de Saint-Exupéry 25bis Avenue Jean Capelle 69621 VILLEURBANNE Tel : 04 72 43 71 70 Fax : 04 72 43 72 37 ed.materiaux@insa-lyon.fr
MEGA	**MECANIQUE, ENERGETIQUE, GENIE CIVIL, ACOUSTIQUE (ED n°162)**	M. Philippe BOISSE Secrétaire : Mériem LABOUNE Adresse : INSA de Lyon École Doctorale MEGA Mériem LABOUNE Bâtiment Antoine de Saint-Exupéry 25bis Avenue Jean Capelle 69621 VILLEURBANNE Tel : 04 72 43 71 70 Fax : 04 72 43 72 37 mega@insa-lyon.fr Site web : http://www.ed-mega.com
ScSo	**ScSo*** M. OBADIA Lionel Sec : Viviane POLSINELLI Insa : J.Y. TOUSSAINT	M. OBADIA Lionel Université Lyon 2 86 rue Pasteur 69365 LYON Cedex 07 Tél : 04.78.69.72.76 Fax : 04.37.28.04.48 Lionel.Obadia@univ-lyon2.fr

*ScSo : Histoire, Geographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Acknowledgments

This work was conducted within ST Microelectronics, Crolles, France, in collaboration with Laboratoire Ampere UMR CNRS 5005, Villeurbanne, France.

First and foremost, I would like to express my gratitude to Philippe Candelier, manager of the Fuse solutions team, for entrusting me with this role. His constant support and devotion to this research project was absolutely crucial in the successful outcome. I particularly appreciated his expertise and his supervision during the past three years.
I am deeply thankful and grateful to Joël Damien, leader of the analog design team, for his contribution and support. He greatly contributed to the present work by his limitless creativity and his outstanding teaching abilities. I will always remember and use the numerous advices and and helpful approaches related to analog circuit design that Joël shared with me.
I do owe a lot to my advisor, Professor Bruno Allard. Bruno taught me, guided me and immensely contributed to the accomplishment of my research work. Working 6 years with him during my Master and Ph.D. experiences was a real pleasure.
Working in the group Fuse Solutions was a great opportunity. I want to thank Elise Le-Roux for designing the antifuse bitcells. I did enjoy breaking down hundreds of them. I also enjoyed our scientific collaboration and her expertise. I also take to opportunity to wish good luck to Thérèse in the Ph.D. journey. I also enjoyed working with Stéphane, Thibault, Laurent, Philippe, Frédéric, Delphine. I want to thank Nouha for her helpful contribution on the layout of the demonstrator. Thank you, Julien, Daniel and Gaetano, from the test department. I hope you guys did not suffer in re-writing the test patterns for the modified demonstrator.
I am grateful to Mustapha Rafik, David Roy and Xavier Federspiel, from the front-end reliability group, for sharing their expertise in reliability. I also want to thank those who helped me in the characterization lab. Thank you Francois, Ruddy, Sébastien and Yannick. I also wish good luck to Laurent who should have now finished his Ph.D. work.
I shared an office with different people that I would like to thank for their help and for making enjoyable my working days. Therefore I am thankful to Gerald, Vincent, Yann, Kamil (good luck for your Ph.D.), Bruno, Noémie and Julien.
I am grateful to Professor Jean-Michel Portal for chairing my Ph.D. committee. I am truly honored to have my Ph.D. dissertation reviewed by Professor Paolo Pavan and Gérard Ghibaudo.
My last acknowledgments go to my family, my wife, my parents, my brother. I truly enjoyed to have everybody attending the Ph.D. defense and for all the encouragements I received from the family.

Abstract

Non-volatile one-time programmable memories are gaining an ever growing interest in embedded electronics. Chip ID, chip configuration or system repairing are among the numerous applications addressed by this type of semiconductor memories. In addition, the antifuse technology enables the storage of secured information with respect to cryptography or else.

The thesis focuses on the understanding of ultrathin gate-oxide breakdown physics that is involved in the programming of antifuse bitcells. The integration of advanced programming and detection schemes is also tackled in this thesis.

The breakdown mechanisms in the dielectric material SiO_2 and high-K under a high electric field were studied. Dedicated experimental setups were needed in order to perform the characterization of antifuse bitcells under the conditions define in memory product. Typical time-to-breakdown values shorter than a micro second were identified. The latter measurements allowed the statistical study and the modeling of dielectric breakdown in a high voltage range, i.e. beyond the conventional range studied in reliability. The model presented in this Ph.D. thesis enables the optimization of the antifuse bitcell sizes according to a targeted mean time-to-breakdown value.

A particular mechanism leading to a high bulk current overshoot occurring during the programming operation was highlighted. The study of this phenomenon was achieved using electrical characterizations and simulations. The triggering of a parasitic P-N-P bipolar transistor localized in the antifuse bitcell appeared as a relevant hypothesis.

The analysis of the impact of the programming conditions on the resulting read current measured under a low voltage was performed using analog test structures. The amplitude of the programming current was controlled in an augmented antifuse bitcell. The programming time is controlled by a programming detection system and a delay.

Finally, these solutions are to be validated using a 1-kb demonstrator yet designed and fabricated in a logic 32-nm CMOS process.

Résumé

Les mémoires non-volatiles programmables une fois sont en plein essor dans le monde de l'électronique embarquée. La traçabilité, la configuration ou encore la réparation de systèmes sur puce avancés font partis des applications adressées par ce type de mémoire. Plus particulièrement, la technologie antifusible présente des propriétés de sécurité autorisant le stockage d'information sensible.

Ce travail de thèse est orienté vers la compréhension des mécanismes de claquage d'oxydes minces sollicités pour la programmation des cellules antifusibles ainsi que l'intégration au niveau système de moyens de détections. Une première étape fut d'étudier les phénomènes de claquage de diélectrique type SiO_2 et à haute permittivité sous l'application d'un fort champ électrique. Des techniques de mesures dédiées ont été développées afin de réaliser des caractérisations dans les conditions de programmation des mémoires antifusible sollicitant des temps au claquage inferieurs à la micro-seconde. Ces mesures ont ensuite permis l'étude statistique du claquage des diélectriques ainsi que la modélisation sous de hautes tensions ; hors des gammes étudiées traditionnellement dans le domaine de la fiabilité. Le modèle proposé permet l'optimisation des dimensions d'une cellule élémentaire en fonction d'un temps au claquage défini au préalable.

Un mécanisme inattendu occasionnant un sur courant substrat a également été mis en évidence pendant la phase de programmation. L'étude de ce phénomène a été réalisée par des caractérisations électriques et des simulations afin de conclure sur l'hypothèse d'un déclenchement d'un transistor bipolaire parasite de type PNP dans la cellule antifusible.

L'impact des conditions de programmation sur le courant de lecture mesuré sous une basse tension a également été analysé. Des structures de tests analogiques dédiés ont été conçues afin de contrôler l'amplitude du courant de programmation. Le contrôle du temps de programmation est quant à lui accompli par un système de détection de courant et de temporisation.

Finalement, ces solutions seront à valider par un démonstrateur d'une capacité de 1-kb conçu et fabriqué sur une technologie CMOS standard avancée 32nm.

Author's bibliography

Journal

M. Deloge, B. Allard, P. Candelier, J. Damiens, E. Le-Roux, M. Rafik, *Application of a TDDB model to the optimization of the programming voltage and dimensions of antifuse bitcells*, IEEE Electron Device Letters, Volume 32, N° 8, pp. 1041-1043, August 2011.

International Conferences

M. Deloge, B. Allard, P. Candelier, J. Damiens, E. Le-Roux, M. Rafik, *Lifetime and wearout current modeling of ultra-thin oxide antifuse bitcells using transient characterization*, IEEE International Memory Workshop, pp. 181-184, 2010.

M. Deloge, B. Allard, P. Candelier, J. Damiens, E. Le-Roux, M. Rafik, *Very fast transient time-to-breakdown measurements on ultra-thin oxide antifuse bitcells using a RF setup*, IEEE 33rd International Convention MIPRO, pp. 30-32, 2010.

M. Deloge, B. Allard, P. Candelier, J. Damiens, E. Le-Roux, M. Rafik, *Understanding the influence of antifuse bitcell dimensions the programming time and energy using an analytical model*, IEEE International Integrated Reliability Workshop, pp. 135-138, 2010.

Contents

CHAPTER 1

Introduction

The emergence of more and more complex System-on-Chips (SoC) leads to an increasing demand for embedded non-volatile memories. The famous flash, albeit dense and fast, is an expensive technology due to a dedicated process. The so-called One-Time Programmable (OTP) memories enables a full-compatibility with core CMOS processes. The low-cost property is therefore appealing for SoC manufacturers.
The present Ph.D. thesis deals with antifuse memories that are used for 10 years as embedded OTP memories for the following applications:

- **Code storage**
- **Secure encryption keys**
- **Analog trimming and calibration**
- **Chip ID**
- **Chip and processor configurability**

Some companies, for instance, Sidense [1] or Kilopass [2] develop almost exclusively non-volatile OTP memory intellectual properties.
In spite of a long period of development and a strong maturity, there is still room for improvement. The useful field programming applications is enabled by a charge-pump circuit that generates the programming voltage on the chip. However, the high voltage amplitude and the energy needed to program the memory leads to a bulky generator. The area occupied by the peripheral circuitry is a heavy burden but essential for the proper operation of this complex system. There is therefore an opportunity in saving a significant circuit area if the programming voltage amplitude and the energy are reduced.

Objective and thesis contents

The prime objective of this PhD is to propose innovative and advanced programming schemes in order to drastically reduce the programming voltage amplitude, the programming current amplitude and the programming time.
To succeed in this assignment, a thorough knowledge of the antifuse bitcell is essential. A large part of the present PhD thesis is focused on the study of the underlying programming physical phenomenon: the breakdown of the ultrathin gate-oxide of a capacitor. A schematic waveform of the current flowing through a capacitor is depicted in figure 1.1. The different steps in the degradation process are highlighted with the corresponding chapters.

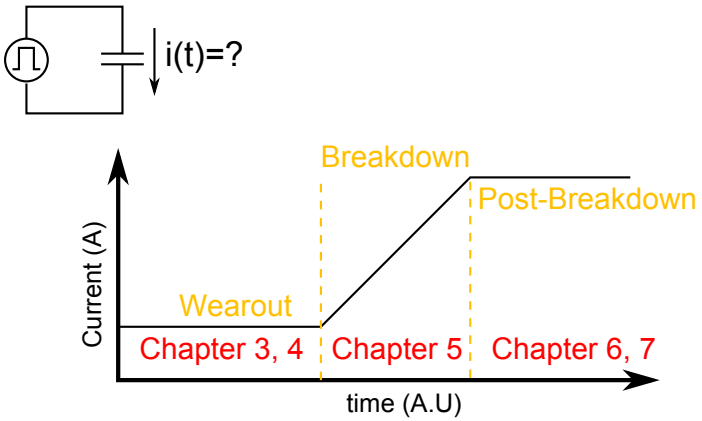


Figure 1.1: Schematic waveform of the capacitor current during a constant high voltage stress.

The semiconductor memory is a significant market in the semiconductor industry. Different non-volatile memory technologies are presented in chapter 2 with a particular focus on One-Time Programmable memories.

A state-of-the-art on the physics of the ultrathin gate-oxide breakdown is detailed in chapter 3. Latest works on the modeling and observations of this failure mode reported in literature are presented and compared to the context of antifuse memories. Recent results on the breakdown of high-K dielectric materials are also presented.

The measurement of the programming sequence of an antifuse bitcell is necessary to investigate the gate-oxide breakdown under a high voltage stress. Experimental setup featuring a high bandwidth are needed for this purpose. Several innovative characterization methodologies are put in place to analyze the gate-oxide breakdown in a high voltage range and short time-to-breakdown.

Chapter 4 deals with the characterization and the modeling of the time-to-breakdown and the wearout current. These two parameters define the first phase of the pro-

gramming sequence. Consequently, it is worth to understand the proportion of the programming time taken by the wearout phase. The second step consists in the optimization of this phase in terms of timing and impact of the circuit area.

The characterizations performed on various antifuse bitcells have revealed an unexpected phenomenon triggered by the breakdown event and leading to a tremendous bulk current overshoot. As a first approach, electrical characterizations were performed. Then, the root cause of the mechanism was investigated. The study is reported in chapter 5.

The post-breakdown phase and its contribution to the read current amplitude is considered in chapter 6. The design of a test structure dedicated to the control of the post-breakdown current amplitude is presented. Thus, the dependence of the read current on the post-breakdown conditions are discussed.

Finally, the design of an advanced antifuse memory demonstrator in a logic 32-nm CMOS process is presented in chapter 7. The bitcells organized in a memory array of 1-kb feature a capacitor with a high-K/metal-gate stack. The functionalities enable an independent control of the post-breakdown time and current amplitude. Consequently, the contribution of the post-breakdown phase on the read current can be studied in a large statistical population of advanced bitcells organized in a dense array, as in a product.

Conclusions on this PhD thesis are drawn in chapter 8 and future perspectives are given.

CHAPTER 2

Non-volatile semiconductor memories

Semiconductor memories are studied and used for more than forty years. In fact, it is pretty difficult to remember how it all started. There is a patent filled in 1957 describing a field programmable memory made of electrical fuses connected to each cross-over point of a matrix [3]. Even though this memory is not "semiconductor", key memory aspects are already emphasized such as the field programmability. This memory is detailed further in this chapter.

First semiconductor memories came out ten years after the invention of the so-called storage matrix. In 1967, Robert Dennard scientist at IBM detailed the invention on the Dynamic Random-Access Memory (DRAM) [4]. Three years after, the Intel 1103 became the first commercially-available 1024-bit DRAM chip. The commercialization of this circuit was one of the major breakthrough in the microelectronics history. Gordon Moore, cofounder and chairman emeritus of Intel said: *"it was the chip that really got Intel over the hump to profitability"* [5]. The success-story of Intel is well known. Nowadays, DRAM is used in every single personal computer, workstation or smartphone.

1980 is also an historical date when Dr. Fujio Masuoka (Toshiba Corp) invented the Flash memory. It was presented at the IEEE 1984 International Electron Device Meetings (IEDM) [6]. There is no need to mention how widely the flash memory is used today.

Semiconductor memories were an essential component in the quest of high performance electronic devices. From this brief historical background, it can be noticed that huge research efforts have been made since the beginning of microelectronics and will be undoubtedly pursued in the future.

This first chapter aims at presenting the semiconductor memory landscape and the place occupied by antifuse memories.

The market share for three years and forecasts of the semiconductor industry and the impact of semiconductor memories on the total IC market are presented in section 2.1. Due to a stringent demand of particular performance for different product applications, many different technologies have been developed. Furthermore, emerging memories are presented and compared.

Section 2.2 is focused on One-Time Programmable (OTP) memories. After an historical background, two eFuse memory technologies: the polyfuse and the metal fuse are presented. Programming mechanisms and state-of-the-art products are reported and compared.

Antifuse memories are detailed in section 2.3. A state-of-the-art of different bitcell architectures provides a better insight into the antifuse memory technology. The performance of late products reported in literature are compared.

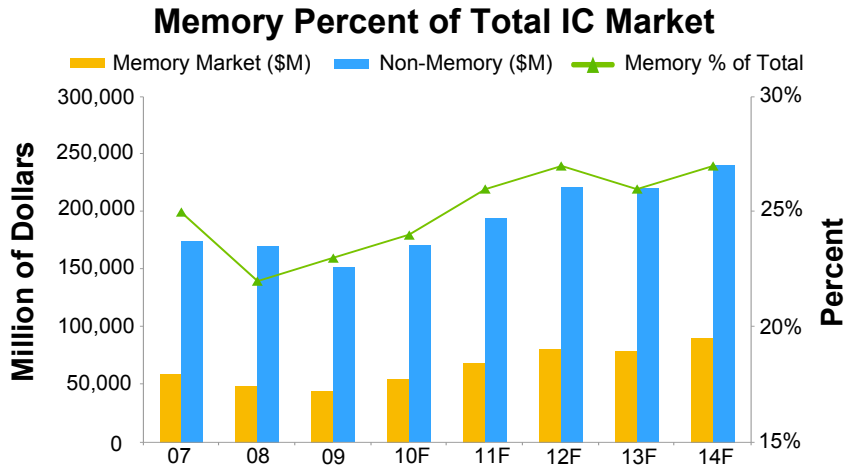
The performance of eFuse and antifuse memories are compared in section 2.4. Pros and cons are listed in order to identify the domain of applications addressed by both technologies.

Conclusions are drawn in section 2.5.

2.1 Types of semiconductor memories

2.1.1 Market trends

Before presenting the different types of semiconductor memory technologies, it is worth to take a look on the market during the past years and the trends for the future.
The worldwide semiconductor market has been growing from the middle of the 60s till nowadays. The market share reached $152 billion in 2009. The memory market has been following this trend and account for approximately 25% of the total semiconductor market. The evolution of the total IC and memory market is shown in figure 2.1.



Memory Market ($M)	59,300	48,946	44,810	55,024	68,835	81,104	79,181	90,785
Semiconductor Market ($M)	174,855	170,474	151,957	170,866	194,496	221,813	220,456	240,640
Memory % of Total	25%	22%	23%	24%	26%	27%	26%	27%

Figure 2.1: Semiconductor memory market and foreseen trends [7]

Despite a negative impact of the financial crisis on the IC market in 2008 and 2009, forecasts are optimistic. Also, the memory percent of total IC market is foreseen to increase up to 27 % in 2014.
The increase of the semiconductor memory market is also an interesting indicator. From 2009 to 2014, the average market growth is 10%.
Taking into account the market indicator, it can be seen that the semiconductor memory is a major part of the global semiconductor market. Furthermore, the increasing demand for evermore complex chips leads to a continued growth for several years.

2.1.2 Classification

In the era of computers, laptops, smartphones and more generally hand-held devices, these technologies would likely not exist without semiconductor memories. There are, in fact, two main categories:

- **Volatile Memories:** a first method of data storage is based on setting a state in a flip-flop bistable circuit. The memories programmed using this approach are known as Static Random Access Memory (SRAM). A second programming method relies on charging a capacitor. However, the charge needs to be periodically refreshed due to the leakage current. Consequently, the memories composed of capacitors are called Dynamic Random Access Memory (DRAM). According to the two programming methods, it can be

understood that the information is lost if the power is turned off. This is the reason why SRAM and DRAM are categorized among volatile memories.

- **Non-Volatile Memories (NVM)** There is obviously a need for a memory able to store data even when the power is turned off. In fact, several technologies which overcome the limitation of SRAM and DRAM are reported and are presented in this chapter.

A non-exhaustive classification of semiconductor memories is shown in figure 2.2.

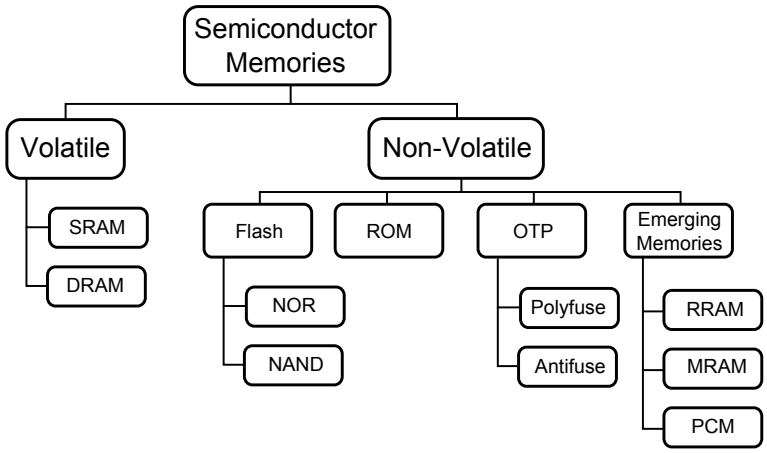


Figure 2.2: Non-exhaustive classification of semiconductor memories.

Before presenting the different types of memory, it is worth to link the diagram shown in figure 2.2 with the market share. 95% of the semiconductor memory market are taken by DRAM and Flash [8]. The remaining 5% are obviously taken by all the other technologies.

2.1.2.1 Stand-alone and embedded memories

Volatile and non-volatile memories are fabricated as stand-alone or embedded devices. The most famous stand-alone memory is undoubtely the DRAM. Every computers or workstations are equipped by DRAM modules as shown in figure 2.3.

Figure 2.3: An advanced Samsung DDR3 DRAM module [9].

In this example, the entire circuit is dedicated to the memory function. Another famous stand-alone memory is the USB smart key in which NAND flash circuits

are assembled. Like for a DRAM, this device is used for data storage. Stand alone memories features usually very high density.
A memory can be also embedded in a System-on-Chip. In the latter case, data storage is not the main function of the circuit. For instance, SRAM are used as cache or data buffer in a microprocessor. In a microprocessor, the SRAM performance has a direct impact on the calculation operation. The SRAM circuits can be easily noticed on the photograph of the Six-Core Intel Xeon 5600 *"Westmere"* die shown in figure 2.4.

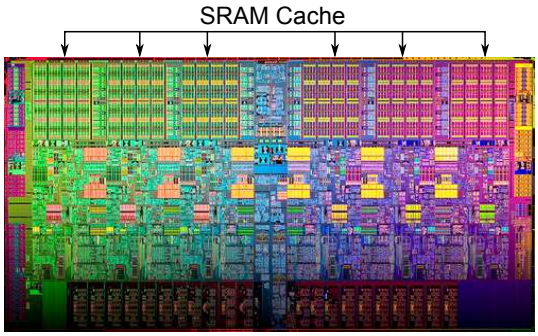


Figure 2.4: Six-Core Intel Xeon 5600 Westmere [10].

The cache memory occupies a large part of the multi-core processor die. In fact, the SRAM is a part of the processor and has a significant impact on the operating speed. Other memory technologies are used as embedded such as flash, ROM and One-Time Programmable (OTP).

2.1.2.2 Memory performance

For providing a glance of the memory performances, the SRAM, antifuse and flash memories are compared using the graph depicted in figure 2.5.

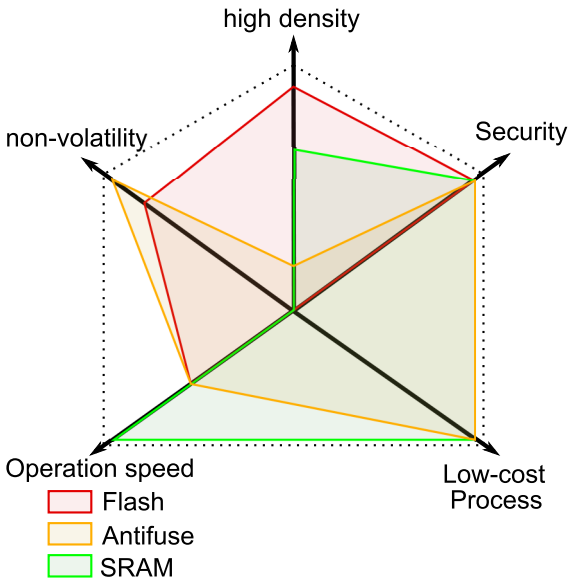


Figure 2.5: Classification of semiconductor memories according to their performance.

Five main properties are taken into account such as high density, security, low-cost process, operation speed and non-volatility. An ultimate memory may encompass these performance by featuring a small cell, an advanced degree of data security, no additional process steps i.e. low-cost, fast operation and finally non-volatility. However, such memory technology does not exist in the semiconductor memory industry.

Different technologies were designed in order to provide appropriate solutions for different applications. As shown in figure 2.5, the SRAM is the fastest memory and is fully compatible with a logic CMOS process. The SRAM is therefore widely used as cache memory in microprocessor (see in figure 2.4). However, the main limitation is the volatility. The flash memory is dense and non-volatile, therefore, appropriate for mass storage. The operation speed and the process cost are the limiting performance. The advantages of antifuse memories are the high level of security, the low-cost property and the non-volatility.

From this simple comparison, it can be understood that different memory technologies are needed for different applications. The markets addressed by the different are therefore separated. Some emerging memories combine advantages of flash and DRAM for example. However, they are in a early stage of development and are not available in production yet.

The different memory technologies are presented along this section. Since this Ph.D. work is focused on antifuse memory, the non-volatile memories are detailed.

2.1.2.3 Memory architecture: random access

A random access memory (RAM) usually refers to volatile memories such as SRAM and DRAM as presented in 2.1.2. However, most NVM are also organized in order to be randomly accessed i.e. any bit of data can be accessed at any time. A simple block diagram of a RAM is depicted in figure 2.6.

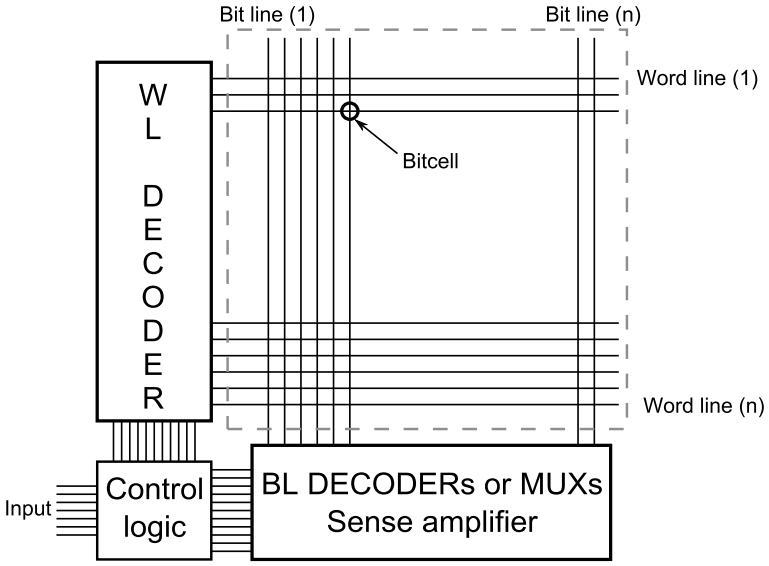


Figure 2.6: Block diagram of a random access memory.

A random access memory is composed of peripheral circuit blocks surrounding an array. This array is accessed using Word Lines (WL) and Bit Lines (BL). A memory bitcell (which can be of any type) is connected at each intersection. For example, an array of 32 WL and 32 BL contains 1024 bitcells. Hence, the memory density is 1024-bits or 1-kb.
To use the memory, the user sets inputs on a control logic block in order to select the bits to be programmed or read in the array. A WL address is decoded, thereby selecting an entire row while every other WLs are unselected. After the WL selection, a single BL or column is accessed at a BL address. Since a single WL and a single BL are selected, only one bitcell is accessed in the entire array. Thus by changing the row and column addresses, any bit can be selected.

2.1.2.4 Read-Only Memory (ROM)

As its name suggests, a Read-Only Memory cannot be reprogrammed by the user. Therefore, data are defined prior to fabrication. The basic operation is illustrated in figure 2.7.

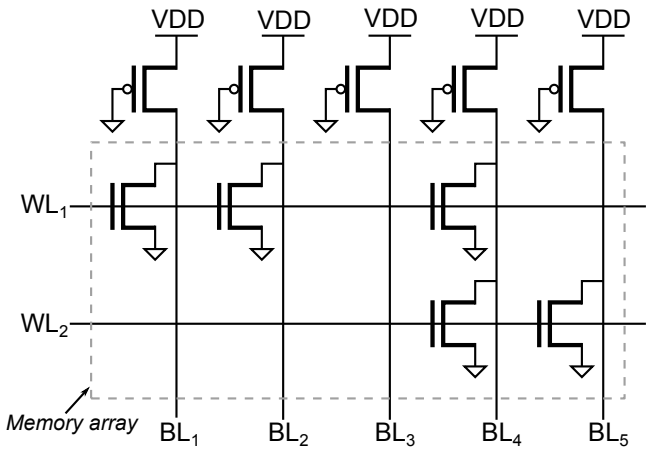


Figure 2.7: An example of ROM [11].

MOSFETs are organized in a memory array and are randomly accessed. The matrix is supplied by VDD through pull-up PMOS transistors. Considering WL_1 set to a high level and WL_2 set to a low level, BLs BL_1, BL_2 and BL_4 are pulled down to ground while BL_3 and BL_5 are pulled up to VDD. Data can be read bit-by-bit by sensing one bit line at a time or word-by-word by sensing every bit lines. Finally a code is stored in the ROM by connecting a MOS transistor at the desired cross-over points of the memory array. The programming operation is performed either using a metal or implant mask. This memory technology is therefore fully compatible with a logic CMOS process or derivative thereof. Furthermore, a ROM features a very high density.

2.1.2.5 Flash Memory

Flash memories have many attractive features. They are non-volatile, electrically programmable, electrically erasable and very dense.
A flash bitcell is composed of a single device. In fact, it is a NMOS transistor featuring an additional floating gate interposed between the channel and the gate-oxide of the MOS transistor. A typical cross-sectional view of a flash bitcell is depicted in figure 2.8.

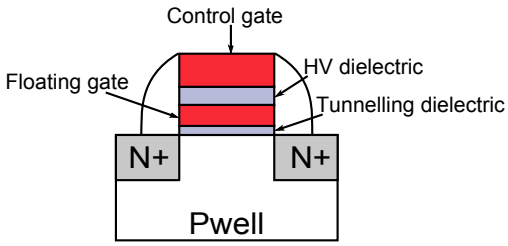


Figure 2.8: Cross-sectionnal view of a FLASH bitcell.

The basic programming operations is performed by trapping charges in the floating gate using a high voltage. Thus the threshold voltage of the MOS transistor is shifted and a data can be read by means of sensing the drain current of the flash bitcell.
The floating gate is strongly isolated and the carriers trapped cannot be discharged for many years (under nominal conditions). To erase the flash bitcell, a high voltage is used to pull off the electrons. Hence, the threshold voltage is shifted back to its nominal value.
A typical characteristic representing the drain current versus the control-gate voltage is shown in figure 2.9.

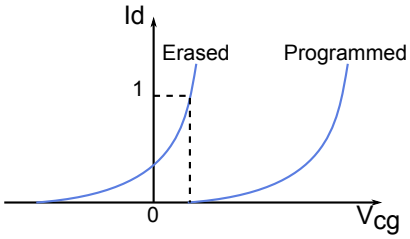
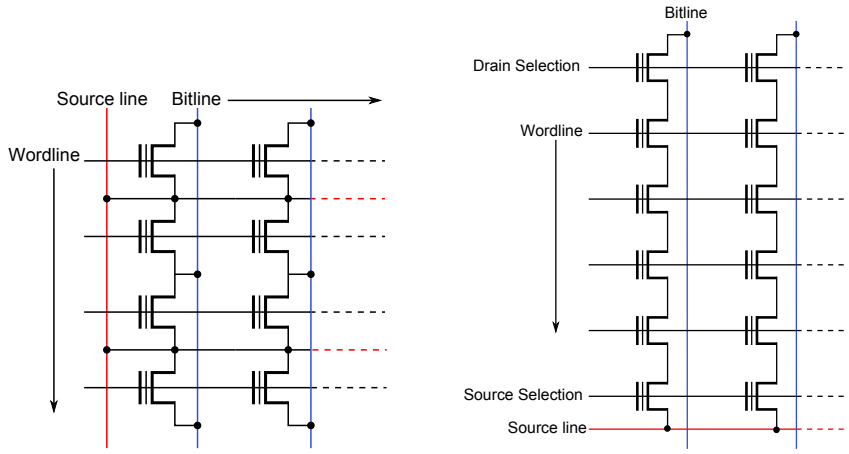


Figure 2.9: Illustration of V_t shift for an erased and programmed flash bitcell.

There are two main architectures of flash memories. In a so-called NOR flash, bitcells are randomly accessed whereas the access is sequential in a NAND flash. In other words, data are accessed in an ordered sequence. The topologies of NOR and NAND architectures are illustrated in figure 2.10.



(a) NOR flash (b) NAND flash

Figure 2.10: Basic schematics of NOR and NAND flash memory architectures.

In the NOR architecture (see in figure 2.10(a)), each drain is connected to a bit line while each source is connected to a source line. Gates are connected to word lines. The access is therefore random.

Bitcells are connected in series in the NAND architecture as illustrated in figure 2.10(b). However, the drain of the top cell and the source of the bottom cell are accessed via a bit line and a source line respectively. Control gates are connected like in the NOR architecture.

To read out a flash bitcell in a line, all the word line except one are pulled up above the threshold voltage of a programmed bit while the other is pulled up above the Vt of an erased bit. The whole group of series cells conducts if the selected bit is not in programmed state.

Due to the series connection of the bitcells, the read access time on a bit line is longer than in a NOR architecture. However, the programming operation performed by block allows a significant reduction of the programming time. Finally, it can be noted that drain contact are not required for each bitcell. The density of a NAND flash is therefore significantly improved (40% compared to a NOR flash) [12].

Both architectures encompass advantages and disadvantages. Consequently there used in different domains of application. The cost per bit of a NAND flash is lower than a NOR. This is the reason why NAND flash are mostly used in systems requiring a high storage capacity. USB smart keys and Solid-State Drive (SSD) are NAND-based flash memories.

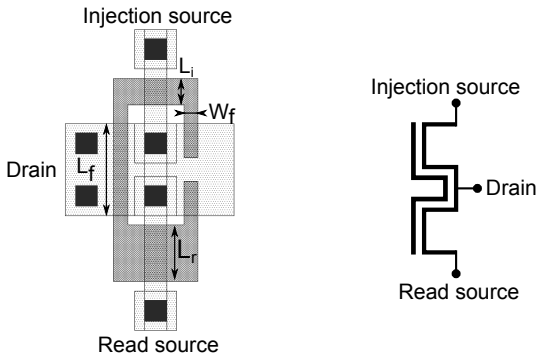
Even though the market was dominated by NOR flash until 2004. The exceptional growth of the hand-held device market lead to the decline of the NOR architecture. However, the advantage of NOR flash is the random access. They are still widely

used as embedded memory in automotive application for example [13].
Although flash memories have very attractive features, the main limitation comes from the complexity i.e. the high cost of the fabrication process [14, 15]. Indeed the floating gate requires additional masks and process steps while the different gate-oxide layers have very particular requirements in order to guarantee the charge retention and to handle the high programming voltage. As a consequence, flash memories are fabricated using an expensive dedicated process. A comparison of the process steps required in a logic CMOS process and in a flash process is given in table 2.1.

Process step	**Logic**	**Flash**
Isolation formation	•	•
High Voltage Wells		2 masks
NVM Array well		1 mask
Tunnel Oxidation		•
Floating Gate Deposition/Patterning		1 mask
ONO/Patterning		1 mask
Low Voltage Wells	•	•
DGO Wells	•	•
High voltage oxidation / Patterning		1 mask
DGO Oxidation /Patterning	•	•
Low Voltage Oxide Growth	•	•
Gate Deposition	•	•
NVM Patterning		1 mask
NVM Source Halo implant		1 mask
NVM Drain Implant		1 mask
Gate Patterning	•	•
High Voltage LDD implants		2 masks
DGO LDD Implants	•	•
S/D Backend Processing	•	•
Masking Step Added		**+11**

Table 2.1: Comparison of required mask steps between a logic and a flash CMOS process [16].

The addition of 11 masks is obviously a limitation. Therefore, flash memories (NOR and NAND) are profitable only if it occupies a wide chip area and or if the memory is a core function.
The complexity of the flash memory fabrication process leads to several prototypes and products of "low-cost" flash e.g. compatible with a logic CMOS process [17–19]. A example of bitcell is depicted in figure 2.11.



(a) Layout (b) Schematic

Figure 2.11: Layout and schematic of the two-channel Y-flash bitcell [18, 19]

The operation principles are similar: charges are trapped in a floating gate. However this floating is the same fabricated for core CMOS devices. In other words, the gate and the dielectric are not optimized for charge retention.
Due to the requirements on the gate-oxide of logic CMOS process, performance are lower than flash memories fabricated using a dedicated process. Furthermore, the constant reduction of the gate-oxide thickness in digital circuits makes the operation of "zero-cost" Many Time Programmable (MTP) memories more and more difficult. A minimum gate-oxide thickness of 70Å is reported in late prototypes [19].
Table 2.2 summarizes the discussion about the three types of flash memory presented in this subsection.

Flash type	**NOR [20]**	**NAND [20]**	**Zero-cost MTP [19]**
Maximum density	1 Gb [21]	256 Gb [21]	64kb
Access mode	Random	Sequential	Random
Access time	100ns	50μs	-
Write time	2Mb/s	+10Mb/s	10ms
Erase time	1ms	2ms	20ms

Table 2.2: Comparison of performance for NOR, NAND and single poly flash memory

2.1.2.6 Emerging memories

Technology nodes after technology nodes, the gate-oxide thickness has been reduced in order to enhance the performance of CMOS transistors. However, this essential improvement has a rather negative impact on semiconductor memories. Indeed, the charge retention in flash memories is more and more difficult. DRAM need to be more refreshed leading to a higher power consumption. Finally the stability of SRAM is compromised by a lower supply voltage.

Memory designers are facing more and more difficulties to follow the pace of device downscaling dictated by the semiconductor roadmaps. Consequently, emerging memory technologies are being actively developed in order to propose an alternative solution for Flash (NOR and NAND) or volatile memories e.g. SRAM and DRAM.

Resistive RAM (RRAM) The underlying concept of RRAM or Conductive Bridging RAM (CBRAM) is to switch a device somewhat similar to a capacitor between a high resistive state (HRS) and a low resistive state (LRS) [22,23]. A typical bitcell comprises a resistance-switching device connected in series to an access transistor. A schematic and the resistance of a cycling test are shown in figure 2.12.

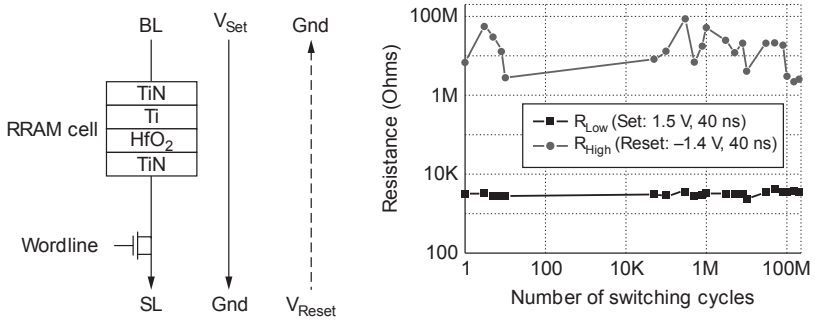
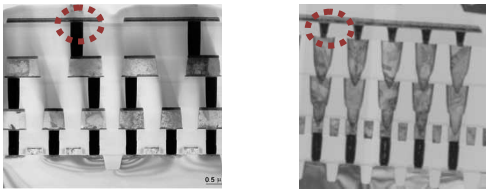


Figure 2.12: Typical RRAM operation and resistance of a cycling test from a high resistive state (HRS) to a low resistive state (LRS) [22].

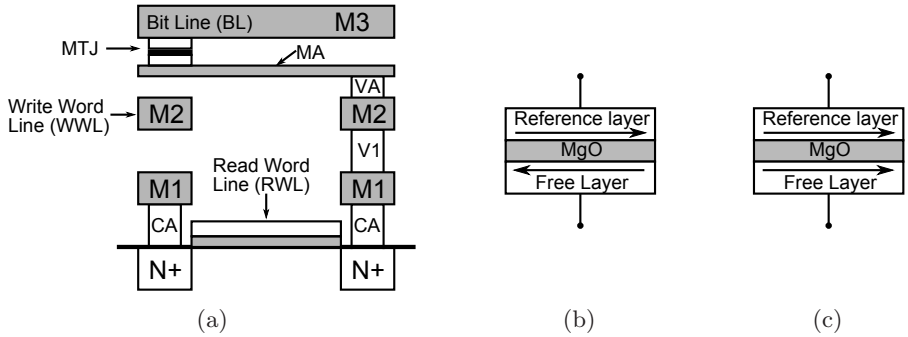
As mentioned previously, the resistance-switching device is a stack comprising a dielectric HfO_2 between two TiN electrodes. By applying a positive or a negative voltage, the capacitor is switched between a LRS and a HRS respectively. Prior to a set or reset cycle, a forming operation is performed under 3.3V. Then, the material is ready to be programmed and erased.
RRAM are still in a early stage of development. However, this technology presents attractive advantages. It is seen in figures 2.13(a), 2.13(b) that storage devices are integrated into the back end of logic CMOS processes. The critical CMOS devices are therefore processed prior to the CBRAM or RRAM bitcells and their performance is not impacted. Moreover, FEOL transistors can be downscaled without impacting the programmable resistor. The materials which compose the resistive switching device are used in conventional CMOS process. Their integration is therefore easier than exotic compound and the risk of contamination is limited. A key issue for the scientists is the understanding of the resistance switching physics [24,25].



(a) (b)

Figure 2.13: CBRAM cell: integration in 180nm logic process (Al BEOL) (b), integration in 130nm logic process (Cu BEOL) (c). Dotted circles indicate the programmable resistor [23].

Magnetic RAM (MRAM) Like a RRAM, a MRAM cell comprises a resistance-switching device connected in series to an access transistor. However, the retention component is a Magnetic Tunnel Junction (MTJ). The MTJ can be switched in an anti-parallel state which corresponds to a high resistance whereas a low resistance is obtained in a parallel state. A schematic of a MRAM cell and an illustration of a MTJ are depicted in figure 2.14.



(a) (b) (c)

Figure 2.14: Cross-sectional view of a MRAM cell (a) [26]. A conceptual view of a MTJ in an anti-parallel state (b) and in a parallel state (c) [27].

A programming operation of a MRAM cell is achieved by means of changing the magnetization of the free layer. In figure 2.14(a), it can be noticed that the Bit Line and the Write Word Line are arranged at a right angle. By passing a current through BL and WWL, an induced magnetic field is generated at the junction and changes its direction. The down scalability of MRAM is made difficult due to the high current required to change the magnetization of the free layer. Furthermore, this technology require specific metal levels as MA, VA and MTJ.
Spin-Transfer Torque (STT) MRAM have recently overcome the heavy burden coming from the high programming current by using a different programming mechanism. The main advantage of the STT-MRAM comes from the fact that the write current scales linearly with device area [28]. A STT-MRAM bitcell is shown in figure 2.15.

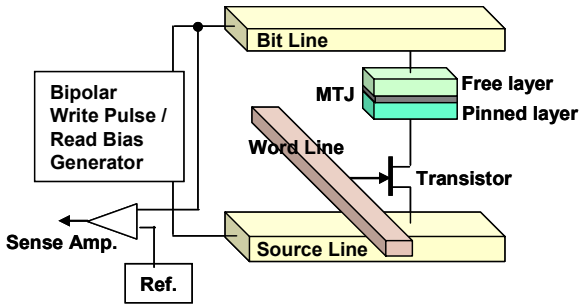
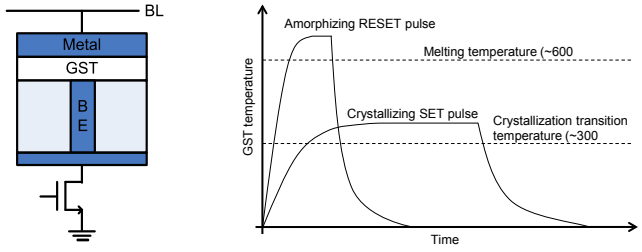


Figure 2.15: Schematic of a STT-MRAM memory cell [29]

The magnetization of the free layer is controlled by the direction of the programming current flowing through the MTJ. A read operation is performed by applying a low negative voltage (-0.1V) on BL while WL is selected. The source line is tied to the ground.
According to authors, STT-MRAM is a potential candidate for the replacement of flash memories and volatile memories due to the combination of a high density and high performance in terms of write speed [30]. Although STT-MRAM is not yet in production, early prototypes has demonstrated the functionality of a 64-Mb memory implemented in 65-nm CMOS technology [31]. The major challenge is the integration of the MTJ in a CMOS process with a reasonable cost.

Phase Change Memory (PCM) The bitcell architecture of a PCM does not differ from RRAM or MRAM. However the resistance-switching material is entirely different. A schematic of a PCM cell and SET and RESET curves are shown in figure 2.16.



(a) (b)

Figure 2.16: Schematic of a PCM cell (a) [32]. Illustration of programming and erase operations [33].

The resistance-switching material is a chalcogenide alloy GeSbTe (GST) which is heated using a bottom electrode BE (see in figure 2.16(a)). The principles of SET (low resistive state) and RESET (high resistive state) are shown in figure 2.16(b).

By applying high and short current pulse, the chalcogenide layer is locally heated and changed into an amorphous material. Consequently, the phase change device is in a high resistive state. The recrystallization involves a lower temperature and a longer pulse.
Scientists from Numonyx have reported a 4-Mb PCM fabricated in 90-nm CMOS technology [34]. The main limitation is the maximum temperature specification of 85Å. However roadmaps indicate that PCM are a very promising candidate for novel applications by combining features from NVM and DRAM [20]. Like STT-MRAM, the cost of the process can be quite high due to the difficult integration of the chalcogenide in a CMOS process.

2.1.3 Comparison

Different types of semiconductor memories were presented throughout this section. For now, volatile memories are used for very fast operation which does not require a permanent storage. Flash memories are able to retain an information for several years without being supplied, however they cannot operate as fast as a DRAM or a SRAM. This is a reason why volatile and non-volatile are used in the semiconductor conductor industry. They do not address the same applications at all.
The continuous downscaling of MOS devices has made difficult the proper operation of traditional memories. Therefore emerging memory technologies are reported. A comparison graph is given in 2.17.

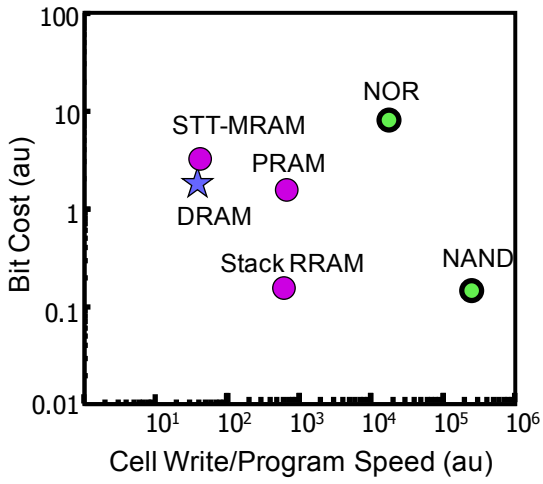


Figure 2.17: Comparison of semiconductor memory technologies (PRAM refers to PCM) [35].

The difference in terms cell write/program speed between DRAM and flash (NOR or NAND) is clearly emphasized. There are 4 or 5 orders of magnitude. The three emerging memory technologies presented in 2.1.2.6 features a write/speed cell

program much better than a flash memory. They are even competing with DRAM and are non-volatile. However, they cannot be considered as an alternative as long as the requirements in terms of reliability and productivity are not fulfilled.

2.2 One-Time Programmable memories

Many different non-volatile memory solutions were presented in section 2.1. Considering their characteristics, it can be noticed that neither of them are compatible with a logic CMOS process. In other words they all require additional process steps and mask levels. In spite of high performance, their expensiveness imply particular domains of application to ensure profitability.
Some applications do not require NVM featuring a lot of rewrite cycles [36]. Applications such as circuit trimming or chip ID require a single but secured programming. One-Time Programmable memories (OTP) are widely used for these purposes. Nevertheless, OTP memories are attractive only if they are low cost i.e. fully compatible with a logic CMOS process. Consequently they can be embedded in System on Chips (SoC) without impacting the performance of the core logic. In addition, derivative processes e.g. imager, BiCMOS or SiGe are also addressed.
OTP memories are also used for circuit repair. Considering a chip with embedded DRAM, extra rows or columns can replace damaged ones by programming a rerouting code in the OTP memory. As a result, the number of usable chips on a wafer is increased [37].
OTP and NVM are also developed internally in the foundries. Consequently, designers can easily implement memory circuits integrated in the SoC. It allows a more flexible circuit configuration and a secure information storage system [38].

2.2.1 OTP in semiconductor markets

The dominating position of DRAM and Flash memories on the semiconductor memory market was emphasized in section 2.1.1. There is, however, a market for embedded OTP memories. The company Kilopass Technology Inc. reports a summary of consumer, mobile, analog and mixed-signal, industrial and automotive markets and the impact of OTP memories in terms of cost savings. The OTP usage in those different markets is summarized in table 2.3.

Market	OTP usage	End product
Consumer	Code storage, ROM patching, Security keys	Set-top boxes
Mobile	Code storage, Security keys, Configuration	CMOS imager sensors, baseband processor
Analog & mixed-signal	Calibration, configuration	Amplifiers, ADCs, DACs
Industrial	Sensors, motor control, video surveillance, environment control, RFID	Microcontrollers
Automotive	Calibration, trimming, code storage, security keys	Control systems, DSPs, in-car communications

Table 2.3: Markets adressed by OTP memories [39]

Considering the mobile market, it can be understood that OTP memories are used in every single mobile phone. The set-top box market is also huge. In the late 2000s, the reception mode of internet, TV and phone signals has radically changed. Now a single box handles multiple complex tasks. OTP are required for different purposes but the storage of security keys is very critical. Even though the market share is much lower than DRAM and flash memories, the application domains addressed by OTP memories put them in a front line in the semiconductor industry. Kilopass reported possible cost savings achieved using OTP memories instead of Electrically Erasable Programmable Memories. Results for different markets are listed in table 2.4.

OTP usage	Before: External EEPROM Total cost/unit	After: Kilopass NVM Total cost/unit	Units / year	Savings for 1 year of production
Consumer: 512kb OTP	$0.25	$0.05	10M	$2M
Mobile: 128kb OTP	$0.10	$0.04	50M	$3M
Analog: 1kb OTP	$0.03	$0.007	70M	$1.61M
Industrial: 8kb OTP	$0.08	$0.02	12M	$0.72M
Automotive: 64kb OTP	$0.08	$0.04	30M	$1.2M

Table 2.4: Costs savings comparison [39]

As introduced in 2.2, OTP or NVM memories fully compatible with a CMOS process allows significant cost reduction compare to an EEPROM. In addition to the compatibility with a logic CMOS process, OTP can also be compatible with derivative process such as imager, RF, automotive, flash ...

2.2.2 Historical background

2.2.2.1 The storage matrix

Regarding the cost savings presented in 2.2.1 allowed by OTP memories, it seems obvious to use a low-cost memory in various applications. In fact the concept of one-time programmability is at least 40 years old. The storage matrix patented in 1957 was one of the first memory and was one-time programmable.
In the 60s was the beginning of equation solving using computers. Scientists needed a technical solution to set numerical constants for various equations at a given time. Finally they required a field programmable memory featuring reliability and flexibility. A schematic of the storage matrix is shown in figure 2.18.

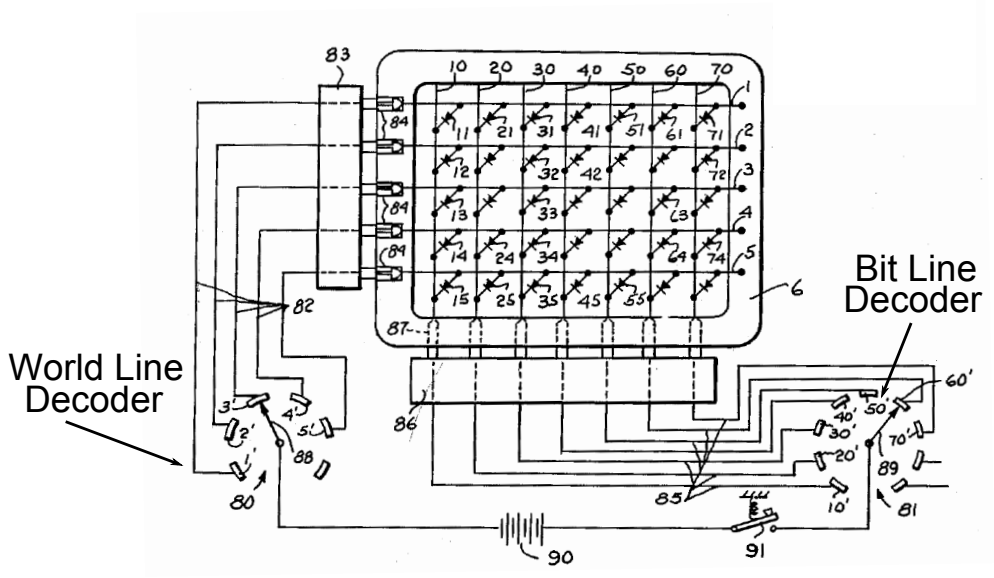


Figure 2.18: Schematic of the storage matrix [3].

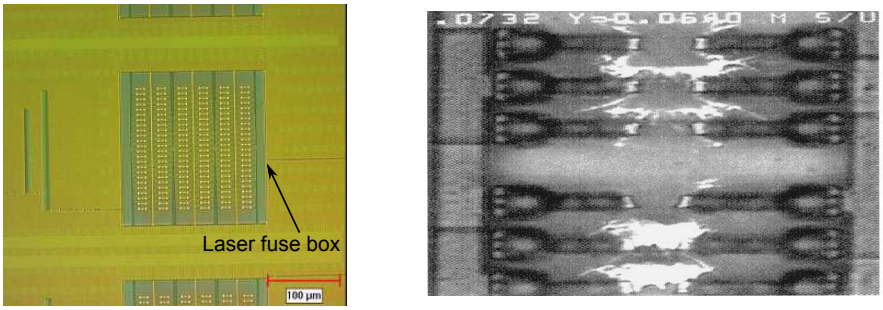
Memory cells are in fact diodes connected at each cross-over point of a matrix. The word line and bit line decoders are rotary switches and allow a random access. In the memory array, word lines and bit lines are connected to each other using diodes. If a diode is removed, the corresponding world line and bit lines are disconnected. Two state are therefore distinguished. In this case, the method for removing diodes is by burning them out electrically. In other word a high current is passed through the device such that the component fails. Hence an open circuit is obtained.

By blowing the desired diodes, a code can be stored in the storage matrix. This code can be read as many time as desired but cannot be reprogrammed. This is the reason why memories operating on this principle are classified in the category OTP or Programmable Read Only Memory (PROM).

2.2.2.2 Laser fuse

A first integrated solution of OTP was reported in the 80s. The so-called laser fuse was mostly used for redundancy with a practical application to VLSI memories [40]. The aim of memory repairing is to improve the manufacturing yield by means of removing defective bits. By burning out laser fuses, the address of a defective bit, row or column is redefined and pointed to a working portion of memory. Even though extra memory bitcells are required, the manufacturing yield is significantly improved [41].

Operation principle A typical laser fuse bitcell is a strip of metal with two electrodes. Like a domestic fuse, a laser fuse is a conducting element. The memory is programmed by blowing the metal strip using a laser, thereby creating an open circuit.
An example of laser fuse box is shown in figure 2.19(a). Fuses burnt using laser pulses are shown 2.19(b). Splash links and craters can be noticed. This is the reason why, the laser fuse technology is not easily scalable. Very small bitcells may lead to programming failure.



(a) (b)

Figure 2.19: Illustration of a 72-bits laser fuse box [42]. Example of programmed laser fuse bitcells [43]

Although the laser fuse technology was used for two decades, limitations can be pointed out.

- **Laser programmability:** Such memory technology is not electrically programmable. As a consequence, additional process steps are required to program the laser fuses using a dedicated machine. Field and embedded programmability are not possible.

- **Non-scalability:** Due to the laser spot size (1.5μm), metal links cannot be scaled down. Since gate length in the most advanced CMOS processes are smaller than 30nm, a bitcell of several μm is not suitable.

- **Additional top level metal layer:** Since laser fuse relies on the back end process, a metal layer on the top is required to be programmed using the laser machine.

Due to these limitations, laser fuses are no longer used in recent CMOS technologies. It can be easily understood that an OTP technology featuring a zero-cost process, scalability and electrical programmability is much more suitable.

2.2.3 eFuse memories

First alternative technologies to laser fuse were reported during the late 90s [44, 45]. Since memories are OTP, the programming operation relies on the impedance variation of an element of the bitcell. Like a laser fuse, a material is destroyed due to the application of a high electrical stress. Two technologies are both widely used: the eFuse and the antifuse. This subsection is focused on eFuse memories, antifuse memories are detailed in section 2.3.

2.2.3.1 eFuse technologies

The programming principle of an eFuse is based on the blowing of a conducting element using a high current. Like a domestic fuse, its resistance is increased by burning out the conducting material. Two technologies are presented in this section.

- **2D Polyfuse:** this solution developed by IBM relies on the front end of the CMOS process. A strip of silicided polysilicon is used as a fuse element [46].
- **3D Metal fuse:** this technology is developed by Intel. The fuse element is located in the back end and is a via between the second and third metal layers [47].

Polyfuse bitcell Examples of a virgin and programmed polyfuse bitcell are shown in figure 2.20.

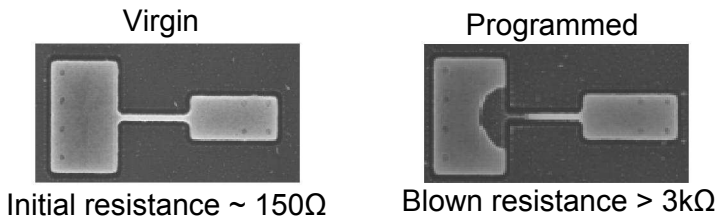
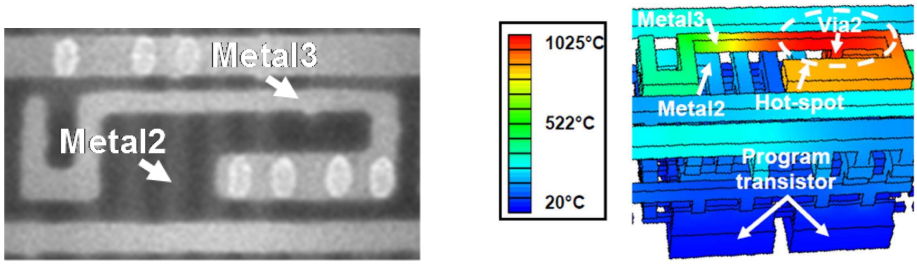


Figure 2.20: Top down SEM views of a virgin and programmed polyfuse [48].

As mentioned previously, the bitcell structure is rather simple. A narrow strip of polysilicon is connected to two electrodes. Since the polyfuse relies on the front end of the CMOS process and more particularly on the polysilicon, no additional process steps are required for the fabrication.
The programming mechanism based on electromigration is clearly illustrated in figure 2.20(b). When a high current is passed through the polyfuse, most of the electrons are transported in the silicided layer due to a resistivity lower than the

polysilicon. The silicide is heated by the high current leading to the electromigration. The depletion is located near the cathode (large electrode in figure 2.20(b)). The electromigration leads to a discontinuous or even disrupted silicided layer. As a consequence, the current is transported by the polysilicon once the polyfuse has been programmed. The resistance of a programmed bitcell is therefore much higher. Even though the specification for the read sensing circuit is 3-kΩ, a programmed resistance higher than 7-MΩ is reported [49].

Metal-fuse bitcell The operation principle of the metal fuse is similar to the polyfuse. There are basically two types of bitcell architecture: a metal strip or an interconnection via between two metal layers [50, 51].
An example of zero-cost metal fuse bitcell developed by Intel is shown in figure 2.21.



(a) (b)

Figure 2.21: (a) 1.37μm^2 bitcell. (b) Three-dimensional of bitcell and electrothermal modeling for element design [47]. OTP memory solution developed by Intel.

Unlike the polyfuse, the metal fuse is in the latter case a 3D structure. The fuse element is stacked above the access transistor. The area occupied by the metal fuse is therefore reduced compared to a planar structure.
The programming mechanism is based on metal electromigration. A local hot-spot is achieved by injecting a high current through the metal fuse until the formation of a consistent void. Therefore, the conducting link between metal 2 and metal 3 is broken.
The metal-fuse bitcell is designed in order to position the hot spot near a weak point of the structure. According to authors, the fact of localizing the melting point in the via minimize the visual exposure. Thus, the metal-fuse technology provides a security protection for product applications.

2.2.3.2 eFuse macros

Many eFuse solutions have been reported for the last 15 years. Most of them were polyfuse memories. However, the most advanced CMOS technologies e.g. 32nm

feature a metal gate instead of a polysilicon gate. Consequently the metal fuse is very trendy. Several eFuse memory products are compared in table 2.5.

Reference	**[44]**	**[52]**	**[53]**	**[46]**	**[47]**
Year	1997	2006	2007	2008	2010
Technology	0.25μm	90-nm	65-nm SOI	45-nm SOI	32-nm
Fuse element	polyfuse	polyfuse	polyfuse	polyfuse	metal fuse
Prog current	20mA	10mA	12mA	7mA	-
Prog voltage	2.5V	-	1.5V	1.2V - 1.8V	1.8V - 2V
Prog time	100ms	200μs	20μs	2μs	1μs
Fuse width	2μm	1.2μm	0.12μm	-	-
Fuse length	0.27μm	0.12μm	0.06μm	-	-
Bitcell area	-	-	6.2μm^2	3.6μm^2	1.37μm^2
Memory density	64-b	-	4-kb (array)	4-kb (array)	4-kb (array)

Table 2.5: Comparison of reported eFuse solutions.

EFuse memories have evolved along the CMOS technology nodes. The programming current has been significantly decreased from 20mA in 0.25μm to 7mA in 45nm SOI. Although the programming current of the metal fuse reported in [47] is not disclosed, several mA can be assumed.

The low programming voltage required to program an eFuse is obviously a strong advantage. The solution reported in [46], the macro is compatible with thin-oxide and thick-oxide supply voltages, 1.2V and 1.8V respectively. Such low-voltage programmability allows an implementation using a single gate-oxide thickness. Besides, no bulky charge-pump circuit is required to generate the programming voltage.

The programming time has been greatly improved as well as the bitcell area. In advanced CMOS, the programming voltage pulse width is about few μs and the bitcell area is few μm^2.

To conclude, the functionality of 4-kb arrays has been demonstrated for polyfuse and metal-fuse memories. The memory density can be greatly increased using several memory banks.

2.3 Antifuse memories

An alternative to polyfuse or metal fuse is the antifuse memory. The programming principle is similar, a device is intentionally stressed until failure in order to change its internal resistance. In an antifuse bitcell this device is a capacitor. Unlike a fuse element, an intact capacitor is an insulator whereas a programmed capacitor exhibits a relative low resistance.

2.3.1 Programming mechanism

As briefly introduced, an antifuse bitcell is programmed by means of breaking down a capacitor. To achieve such a failure, the dielectric is stressed under a high voltage. Hence, defects are generated in the gate-oxide until the breakdown event. The breakdown mechanism is illustrated in figure 2.22 by the so-called percolation model [54, 55].

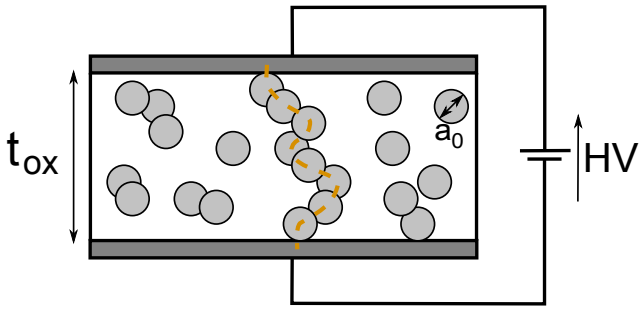


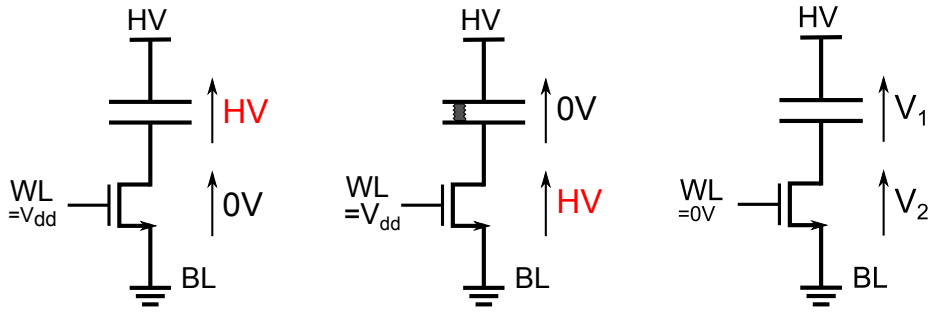
Figure 2.22: Illustration of the percolation model. t_{ox} = gate-oxide thickness, a_0 = defect diameter.

Assuming that a number of defects are randomly generated within the dielectric, the gate-oxide breakdown is achieved by the formation of a contiguous path connecting the two electrodes. This is a statistical approach of the gate-oxide breakdown. Different modeling approaches and a phenomenological study of the breakdown physics are discussed in chapter 2. Only basics are presented in this section for providing insights into the programming operation of antifuse memories.

The time-to-breakdown of a capacitor is dictated by the gate-oxide thickness and the programming voltage as the thinner the gate-oxide, the higher the programming voltage, the shorter the resulting time-to-breakdown. In a CMOS technology, the thin gate-oxide is dimensioned in regarding the reliability in order to sustain a nominal voltage without failing for ten years. As consequence, a much high programming voltage is applied to antifuse bitcells in order to enable a short programming time.

2.3.2 Bitcell architecture

The high voltage required to program an antifuse bitcell in a sufficiently short time leads to particular architectures. A conventional antifuse bitcell topology comprises a capacitor and an access transistor as depicted in figure 2.23 for different voltage operations.



(a) Virgin (b) Programmed (c) unselected

Figure 2.23: Voltage operation in an ideal virgin, programmed and unselected antifuse bitcell.

To program a virgin antifuse bitcell, a high voltage is applied to the HV node while the access transistor is turned on (BL=gnd and WL=V_{dd}). The gate-oxide is therefore being stressed by a high electric field. The drain-to-source voltage of the access transistor is neglected in this example (see in figure 2.23(a)).
Once the short circuit is created in the dielectric, the high voltage is applied to the drain of the access transistor. The voltage operation of an unselected bitcell depends on the state of the antifuse capacitor. A programmed device leads to a high voltage across the access transistor as depicted in figure 2.23(b) whereas the voltages V_1 and V_2 should be distributed such that a virgin capacitor is not stressed while the bitcell is not selected. This operating point in an off mode is set by the difference is leakage current between the capacitor and the turned-off access transistor.
One of the design challenge in an antifuse memory is to cope with the high voltage using an access device compatible with a logic CMOS process. Several bitcell architectures are reported in literature and are presented in the following sections.

2.3.2.1 Drift antifuse bitcell

The drift antifuse bitcell comprises a standard thin-oxide capacitor and a so-called thick-oxide drift access transistor [56]. Schematic and cross-sectional views are depicted in figure 2.24.

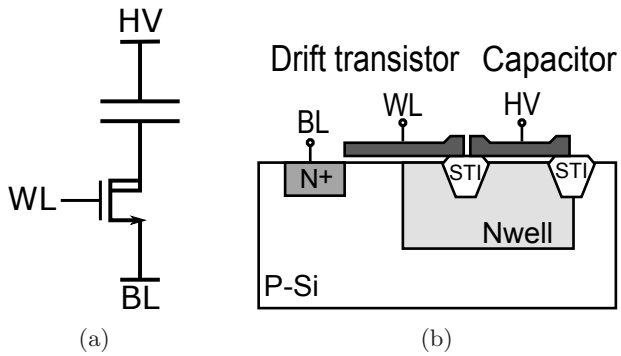


Figure 2.24: Schematic and cross section of a drift antifuse bitcell.

The nominal voltage operation of the drift transistor are extended due to the replacement of a conventional N+/Psub drain junction by a Nwell implantation. Indeed the internal resistance beneath the Shallow Trench Isolation (STI) allows a reduction of the voltage applied on the drain. Thus, the drift transistor sustains a voltage compliant with a the antifuse programming voltage.
The antifuse capacitor is also optimized. A N+ polysilicon gate is chosen instead of a P+ in order to reduce the capacitor breakdown voltage. For a positive stress voltage, the resulting electric field applied across the gate-oxide is higher due to appropriate work functions of the N+ polysilicon and the Nwell.
To conclude, it is important to note that neither additional mask nor process step are required to fabricate the drift bitcell. This architecture is still used within STMicroelectronics and is further discussed in chapter 2.

2.3.2.2 Cascode antifuse bitcell

Cascoding is a well-known design technique in high voltage circuit. Like a drift access transistor, the purpose of a cascode totem is to sustain a high voltage.

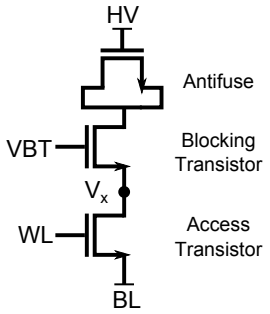


Figure 2.25: Schematic of a 3-transistor antifuse cascode bitcell.

The cascode bitcell comprises a NMOS antifuse capacitor, a blocking transistor to

prevent a high voltage stress and an access transistor [57]. The dimensioning of the cascode and the blocking voltage amplitude V_{BT} avoid the node V_x to reach a voltage higher than the nominal condition.
A thin-oxide capacitor is used. The cascode transistors can be either thick-oxide or thin-oxide.

2.3.2.3 Dual-port cascode bitcell

The antifuse bitcell developed by Toshiba has a particular topology [58, 59]. A thin-oxide PMOS transistor is used as an antifuse capacitor while two different thick-oxide cascode totems are used as write-port and read-port. A schematic is depicted in figure 2.26.

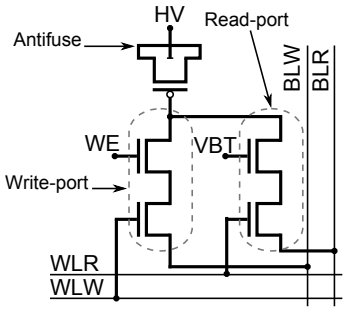


Figure 2.26: Schematic of the dual-port antifuse bitcell

Implementing two ports in the antifuse bitcell allows the simultaneous enhancement of programming and reading properties.
The read port can be optimized in order to reduce the bitline parasitic capacitance. In other words transistors should be quite small. On the other hand, the write port should enable a programming current sufficiently high to lower the broken capacitor resistance. The transistor widths are therefore wide.

2.3.2.4 Multi Antifuse cascode bitell

An antifuse bitcell based on DRAM cell capacitors is developed by Samsung [60]. An illustration is depicted in figure 2.27.

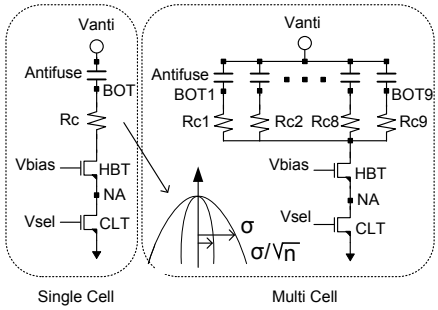


Figure 2.27: Comparison of Single Cell and Multi cell structure with modeling of the interfacing contacts.

The particularity of this bitcell emphasizes an advantage of antifuse memory. Since neither particular process steps nor options are required to fabricate an antifuse capacitor, this technology can be used in derivative processes such as imager, or RF. In this example, authors take advantage of the small dimensions of a DRAM capacitor.
Authors mentioned a limitation of the DRAM capacitor as the high contact resistor which may affect the programmability of antifuse cells. To overcome this limitation, multiple capacitors are used in a single bitcell. Due to the small size of a DRAM capacitor, the area overhead is relatively small. A bitcell is read as programmed if one capacitor is completely broken down. Therefore, a multi-cell antifuse capacitor reduces the standard deviation σ by $1/\sqrt{\text{n}}$, where n is the number of cells.

2.3.3 Antifuse macros

The different macros designed using the bitcells presented in 2.3.2 are compared in table 2.6.

Reference	**[56]**	**[57]**	**[61]**	**[58]**	**[59]**	**[60]**
Year	2000	2006	2006	2007	2009	2010
Technology	0.25-μm	0.18-μm	0.18-μm	65-nm	40-nm	50-nm DRAM
t_{ox}	50Å	40Å	-	19Å	17Å	-
Bitcell area	-	4.2μm^2	-	15.3μm^2	8.1μm^2	-
Prog Voltage	10.5V	6.5V	6V	6.5V	6.5V	6V
Prog current	-	600μA	-	-	-	564μA
Prog time	50ms	20ms	50μs	10μs	10μs	-
Density	-	32kb	1kb-1Mb	1kb-8kb	512b-16kb	-

Table 2.6: Comparison of reported antifuse solutions.

There were significant evolutions in terms of programming voltage and programming time for the last ten years. The reduction of gate-oxide thickness technology node

after technology node has a very positive impact on the performance of antifuse memory. The gate-oxide breakdown plays obviously a major role in antifuse memory programming. The understanding of the mechanism is widely discussed in the chapter and is one of the main topic addressed in this Ph.D. work.

2.4 eFuse versus Antifuse

Both memory technologies are widely used in the microelectronic industry. As a matter of fact, the different OTP technologies presented in this chapter are developed by major semiconductor companies as listed in table 2.7.

Antifuse	**eFuse**
STMicroelectronics	Intel
Samsung	IBM
Toshiba	TSMC

Table 2.7: OTP solutions developed by major semiconductor companies.

There are pros and cons for both technologies. A comparison of the eFuse and antifuse properties is summarized in table 2.8.

	Antifuse	**eFuse**
Process compatibility	++	++
Field programmability	++	++
Prog voltage	-	+
Prog current	+	-
Prog time	-	+
Cell area	-	+
Complexity	-	+
Process variation	+	-
Security	++	–

Table 2.8: Comparison of antifuse and eFuse properties.

The compatibility with a standard CMOS process is crucial for OTP memories. Polyfuse, metal fuse and antifuse comply with this requirement. Another major advantage is the electrical programmability. Memories can be programmed in the field after dicing and packaging.
The main disadvantage of antifuse memory is the high programming voltage amplitude. This voltage is generated using a charge-pump circuit in which capacitors are switched. Thus, the voltage is elevated stage after stage up to six or seven times higher than a nominal supply voltage. It obviously brings reliability concerns. Even

though a charge-pump circuit is required for embedded programming, the memory area is strongly affected by the bulky capacitors. Also, the design of a capacitive DC/DC converter is quite complex. A polyfuse or a metal fuse can be programmed using a regular supply voltage. However, it requires a high programming current (e.g. 10mA) whereas less than a milliamp is enough to breakdown the dielectric of an antifuse capacitor. Therefore it leads to a wide access transistor and back end of line metal tracks to handle such a high current.

Although the antifuse technology is lagging behind polyfuse in terms of programming time, the gap is getting narrower as long as the dielectric thickness in standard MOS device is reduced.

The cell area is also an important parameter for a memory. Since the peripheral circuitry has to be considered for both technologies, it is worth to compare the area per bit instead of the strict bitcell dimensions. A practical example is shown in figure 2.28.

(a) polyfuse (b) Antifuse

Figure 2.28: Floorplan of a 4-kb polyfuse array [46] and of a 16-kb antifuse memory [59]

The area per bit of the polyfuse memory shown in figure 2.28(a) is $9.3\mu m^2$ whereas this parameter is $14\mu m^2$ for the antifuse macro from figure 2.28(b). This property is directly correlated with the circuit complexity of the memory. The architecture of a polyfuse memory is rather conventional. Indeed, the memory core is surrounded by a programming circuitry, senses and decoder. The voltage generator of the programming voltage as well as the bandgap reference circuit occupies a significant circuit area and bring design complexity. The complexity is further increased if a antifuse cascode bitcell is used. In the latter case, another charge-pump circuit is required to generate the gate voltage. This constraint is avoided by implementing a drift bitcell. However, the Nwell involves larger space between bitcells in the

memory array. There is, in fact, a trade-off between the density and the design complexity.

Due to the importance of the thin gate-oxide in a CMOS technology, this parameter is controlled very accurately even from one manufacture to another. The variation of the silicide process can be up to 30%. Although an antifuse memory is more complex than an eFuse in terms of system design, the qualification of the polyfuse bitcell involve a thorough study of the electromigration mechanism to overcome the process variation.

The property which make an obvious distinction between this two technologies is the security. The electromigration mechanism involved in the programming of a polyfuse is clearly visible. The photographs of an intact and programmed bitcell shown in section 2.2.3.1 figure 2.20 demonstrated that it is easy to read out code in a polyfuse array.

The small nanometer size of a gate-oxide breakdown spot is a very attractive property. As a matter of fact, it is difficult to locate the spot using cross-sectional or top bitcell views. A cross-sectional view of a virgin and a programmed antifuse bitcell from Kilopass is shown in figure 2.29.

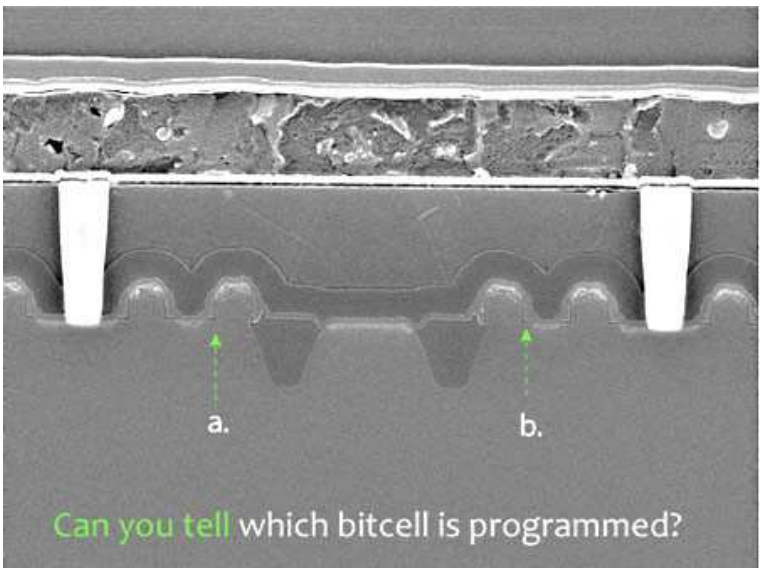


Figure 2.29: SEM cross sectional view of a programmed and virgin antifuse bitcells [62].

Antifuse memories are also robust against passive, semi-invasive and invasive methods. The determination of a word pattern using read current profiles is ineffective because the read current is much lower than the supply current of the peripheral circuitry. Backside attacks, chemical etching or mechanical polishing are unsuccessful because of the thinness of the dielectric and the connection of bitcells in a cross-point array [62].

2.5 Conclusion

Semiconductor memories are strategic products in the semiconductor industry. Even though the market is dominated by DRAM and flash memories in terms of sales, there is an ever-growing demand for different memory technologies in order to address a variety of application domains with specific constraints.
OTP memories are embedded in state-of-the-art SoC for more than ten years. The high degree of security of antifuse memories allows to address product applications with stringent requirements in a standard CMOS process. In spite of a higher level of design complexity and performances slightly lower than eFuse memories, antifuse memories are the best solution for short-term product applications requiring security.
Since an antifuse bitcell is programmed by breaking down the dielectric of a capacitor, the thorough study of this physical phenomenon is obviously essential in order to perform optimization. Furthermore, the CMOS technology node 32nm/22nm features a high-K dielectric and a metal gate whereas devices in previous technologies were entirely made of silicon. Characterizing the breakdown of high-K dielectric is then also essential.
Methods of characterization are needed to study the dielectric breakdown under high voltage programming conditions. Innovative solutions at system level can be designed only if the underlying programming physical phenomenon is practically characterized and eventually modeled.

CHAPTER 3

Antifuse memories and gate-oxide breakdown

An antifuse memory is a complex system. As presented in chapter 2, approximately half of the memory area is occupied by the peripheral circuitry. One of the bulkiest circuit is the charge pump which generates the high programming voltage. Consequently, studying and understanding the influence of the programming conditions on the effectiveness of the gate-oxide breakdown is valuable to optimize the specifications of the aforementioned peripheral circuitry. It is worth to characterize, for example, the impact of the programming voltage and the programming current on the dielectric breakdown.
The high voltage amplitude required to program antifuse bitcells in a short time leads to a particular approach of the gate-oxide breakdown physics. Even though the failure of ultrathin dielectric is studied in reliability, the voltage and time-to-breakdown (T_{BD}) ranges are different than the programming conditions of antifuse memories. Furthermore, advanced antifuse capacitors feature an area much smaller than conventional test structures used in reliability.
The goal of the present chapter is to cover the gate-oxide breakdown physics. Models identified in literature are used to provide a description of the physical phenomenon that focuses the specific conditions associated to the antifuse memory bitcell.
A modeling approach is proposed in section 3.1. The aim is to identify the relevant parameters necessary to characterize and model the gate-oxide breakdown during the programming operation of an antifuse bitcell.

The dielectric breakdown is thoroughly studied for decades regarding reliability (insuring the dielectric robustness in a nominal voltage). Reported models and observations are summarized in section 3.2. Then, conclusions and perspectives are

drawn regarding the validity and the applicability of the different theories for the programming conditions of antifuse memories (insuring the dielectric breakdown for a minimum high voltage).

The process of state-of-the-art CMOS technologies e.g. 32nm, 28nm has changed due to the integration of a high-K dielectric and a metal gate. The impact on antifuse memories is discussed in section 3.3.

Section 3.4 introduces the breakdown modes in reliability analysis and characterization techniques.

Conclusions are drawn in section 3.5.

3.1 Modeling approach

The programming operation of an antifuse bitcell has been briefly introduced in chapter 2, section 2.3. The gate-oxide of a capacitor is broken down by applying a high voltage, i.e., a high electric field across the dielectric. An approach of the programming current waveform is proposed figure 3.1.

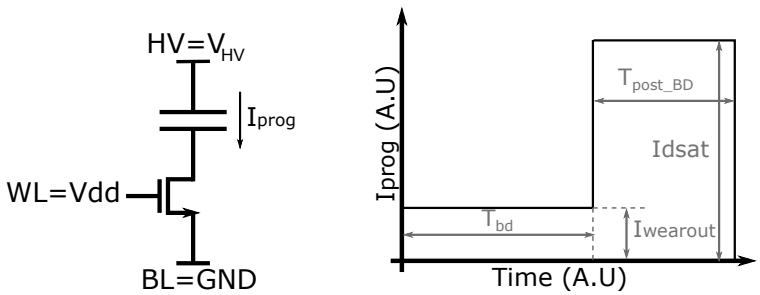


Figure 3.1: Simplified programming operation of an antifuse bitcell.

Appropriate signals are applied to the antifuse bitcell in order to perform a programming operation. The access transistor is turned on while the HV node is pulled up to a high voltage. The gate-oxide is therefore being stressed by a high electric field.
The gate-oxide breakdown physical phenomenon can be divided in three phases regarding the evolution of the programming current:

1. **Phase 1:** a so-called *wearout current* flows through the bitcell dielectric as a result of conduction mechanisms.

2. **Phase 2:** after a duration called *time-to-breakdown* (T_{BD}), the breakdown event occurs. The programming current increases from the wearout current level to the saturation current level of the access transistor.

3. **Phase 3:** The programming current remains steady during the post-breakdown phase defined by the duration $T_{post-BD}$.

The measurement of the programming current, as the schematic waveform shown in figure 3.1(b), allows the characterization of the gate-oxide breakdown. First, knowing how much time is needed to program a bitcell for a given high-voltage amplitude is obviously a key parameter in order to define a programming voltage-pulse duration. This parameter must be obviously minimized in order to ensure a short programming time of the antifuse memory. Second, the contribution of the wearout current on the programming time is also an important parameter. Finally, understanding the impact of the energy consumed during the post-breakdown phase on the resistivity of the breakdown spot is obviously valuable in order to optimize the performance on an antifuse memory. This latter point covered in chapter 6.
Many different challenges are involved in the optimization of the antifuse memory programming operation. The aim of this chapter is to summarize the state-of-the-art ultrathin gate-oxide breakdown modeling approach and characterization methods. The approach presented in this chapter is, however, antagonistic because the programming operation of an antifuse bitcell relies on the failure of a dielectric designed as robust and reliable. Consequently, the relevancy and the applicability of the models and results pointed out in literature are discussed regarding the particular programming conditions of antifuse memories.

3.2 Gate-oxide breakdown mechanisms

The gate-oxide breakdown is the final step of a degradation process. The evolution of the wearout current shown in figure 3.6(b) exhibits a slight increase before the breakdown event. Literature assumes that the gate-oxide is damaged by the tunneling carriers until a severe failure. This approach is illustrated in the diagram depicted in figure 3.2.

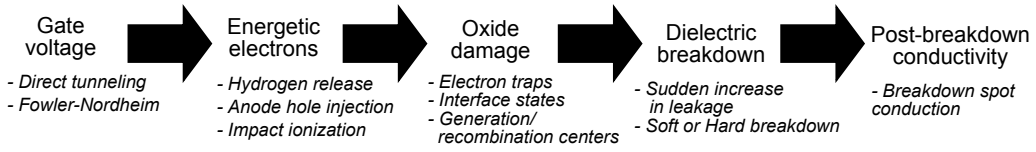


Figure 3.2: Defect generation mechanisms leading to dielectric breakdown [63].

The application of a voltage across a capacitor allows a current to flow through the dielectric. Different current transport processes can be involved such as direct tunneling for a low voltage or Fowler-Nordheim for a high voltage. A consequence of the flow of carriers is the triggering of various physical mechanisms (hydrogen species release, impact ionization ...). Thus, a number of defects are created within the dielectric or at the interface. The accumulation of these defects leads to a conductive breakdown path. There are several criteria that define the breakdown such as hard and progressive. Finally, the breakdown spot should feature a low resistance. The post-breakdown phase is not usually studied in reliability as the studies are focused the event of failure. However, the consequences of this failure are essential in the proper operation of an antifuse bitcell.
The mechanisms presented in figure 3.2 are discussed in this chapter expect the post-breakdown phase which is detailed in chapter 6.

3.2.1 Current transport processes in dielectrics

In spite of the insulating property of a MOS capacitor, the application of a voltage across the dielectric leads to a leakage current. This phenomenon is amplified by the reduction of the gate-oxide thickness.
Tunneling is a common conduction mechanism. A ballistic carrier transport occurs under a low or high electric field by crossing a potential barrier. Hence the tunnel emission depends strongly on the bias condition. Two main mechanisms can be identified as Direct Tunneling and Fowler-Nordheim Tunneling [64]. These two current transport processes are illustrated in figure 3.3.

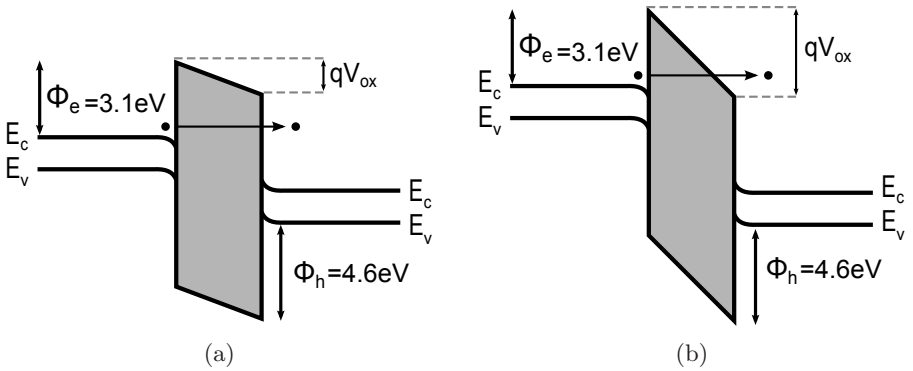


Figure 3.3: Energy-band diagrams illustrating (a) Direct tunneling and (b) Fowler-Nordheim tunneling

In both conduction mechanisms, the current transport is limited by the barrier height of the electrodes, i.e. Φ_e for electrons and Φ_h for holes. In this case, the conduction of electron is dominating due to a lower barrier height.

Either Direct or Fowler-Nordheim tunneling occurs according to V_{ox} amplitude. For a low V_{ox}, carriers crosses a trapezoidal barrier. However, if the capacitor is biased by a sufficiently high voltage, electrons can cross the conduction band. The barrier is hence triangular.
In fact, a direct tunneling current is mostly limited by the gate-oxide thickness whereas the bias voltage, i.e. the triangle barrier shape, is the most impacting factor in Fowler-Nordheim. This two statements are further explained by studying the modeling equation of the two current transport processes.
In the context of antifuse memories, it can be assumed that this two mechanisms may occur in different operating modes. Since a bitcell is read under a low voltage, a direct tunneling current can be expected through a virgin capacitor. On the other hand, a Fowler-Nordheim current transport process may be a good candidate to model the wearout current flowing through the gate-oxide during a programming operation due to the high voltage.

3.2.1.1 Fowler-Nordheim tunneling current

In a Fowler-Nordheim conduction mode, the current density J_{FN} varies with respect to the electric field E_{ox} as derived in equation (3.1) [65]:

$$J_{FN} = C \cdot E_{ox}^2 \cdot \exp\left(-\frac{D}{E_{ox}}\right) \tag{3.1}$$

Where the so-called Fowler-Nordheim parameters C and D are:

$$C = \frac{q^3}{8\pi^2\hbar\Phi_b} \cdot \frac{m_{Si}^*}{m_{ox}^*} \quad \text{and} \quad D = \frac{4}{3}\frac{\sqrt{2m_{ox}^*}}{q\hbar} \cdot \Phi_b^{\frac{3}{2}} \tag{3.2}$$

q is the elementary charge, m_{Si}^* is the electron rest mass, m_{ox}^* is the electron effective mass within the dielectric, $\hbar$ is the reduced Planck Constant ($h/2\pi$) and Φ_b is the injecting electrode barrier height.
As mentioned previously, the current density depends on the electric field E_{ox} and the dielectric material characterized by Φ_b and the electron effective mass m_{ox}^*. It can be noted that the current density is temperature independent.
In order to extract the parameters C and D, (3.1) can be transformed as:

$$\ln\left(\frac{J_{FN}}{E_{ox}^2}\right) = \ln(C) - \frac{D}{E_{ox}} \tag{3.3}$$

Assuming a Fowler-Nordheim conduction mode, (3.3) demonstrates that plotting $\ln\left(\frac{J_{FN}}{E_{ox}^2}\right)$ versus $\frac{1}{E_{ox}}$ yields a slope with a gradient -D and a y-intercept ln(C). Thus, the parameters are easily extracted and a model of a Fowler-Nordheim current can be plotted using (3.1). An application is given in chapter 4.

3.2.1.2 Direct tunneling current

The expression of a direct tunneling current is more complex than a Fowler-Nordheim conduction. An equation based on the Fowler-Nordheim parameters introduced previously is [66]:

$$J_{DT} = \frac{A \cdot E_{ox}^2}{\left[1 - \sqrt{\left(\frac{\Phi_b + qE_{ox}T_{ox}}{\Phi_b}\right)}\right]^2} \times \exp\left[-\frac{B}{E_{ox}} \frac{\Phi_b^{3/2} - (\Phi_b - qE_{ox}T_{ox})^{3/2}}{\Phi_b^{3/2}}\right] \quad (3.4)$$

C and D are the same parameters as defined in expressions (3.2). In a direct tunneling conduction mode, the current is also impacted by the oxide thickness in contrary to the Fowler-Nordheim mode.

3.2.1.3 Frenkel-Poole transport

The Frenkel-Poole transport is due to the emission of trapped electrons in the conduction band. In other words, a carrier moves from one trap to another by thermal excitation. In this case, the conduction is not limited by the barrier height of the electrodes but by the barrier height of a trap. Considering that the gate-oxide is broken due to the accumulation of defects in the dielectric, this type of conduction may occur during the read operation of a programmed bitcell. The Frenkel-poole conduction mechanism is illustrated in figure 3.4.

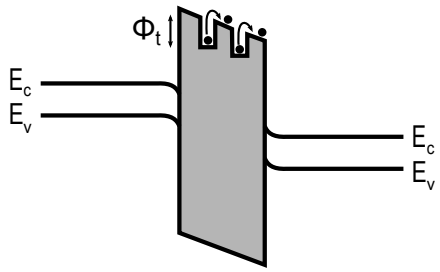


Figure 3.4: Energy-band diagram illustrating the Frenkel-Poole transport.

An expression of the current density J_{FP} is derived in the following equation [64]:

$$J_{FP} = C \cdot E_{ox} \cdot \exp\left(-\frac{q\Phi_t}{kT}\right) \cdot \exp\left(\frac{\beta_{FP}\sqrt{E_{ox}}}{kT}\right) \quad (3.5)$$

C is a constant proportional to the defect density and β_{FP} is the Frenkel-Poole factor defined as:

$$\beta_{FP} = \sqrt{\frac{q^3}{\pi\varepsilon_{ox}}} \quad (3.6)$$

As explained previously, the current density depends on the trap barrier height Φ_p and the electric field E_{ox}. ε_{ox} is the dielectric permittivity. Furthermore, the Frenkel-Poole emission is thermally activated as reflected by the term kT where k is the Boltzmann constant and T the temperature. The Frenkel-Poole occurs mostly at high temperature and high fields [64].
Equation (3.5) can be transformed as:

$$\ln\left(\frac{J_{FP}}{E_{ox}}\right) = \ln(C) + \frac{q\Phi_t\beta_{FP}}{kT}\sqrt{E_{ox}} \tag{3.7}$$

The parameters required to model J_{FP} using (3.5) can be identified by plotting $\ln\left(\frac{J_{FP}}{E_{ox}}\right)$ versus $\sqrt{E_{ox}}$.
The defect density in ultrathin gate oxide is generally very low due to an accurate control of the process. The threshold voltage of MOS transistor is directly affected by those defects. Consequently, direct or Fowler-Nordheim tunneling mainly occurs. The nature of current transport processes in antifuse bitcells is further discussed in chapter 4.

3.2.2 Statistical approach

The gate-oxide breakdown should not be studied on a single device but on a set of devices in order to analyze the statistical dispersion of the time-to-breakdown.

3.2.2.1 Weibull distribution

The Weibull distribution is commonly used in reliability to model a failure mode in a system. The gate-oxide breakdown is obviously a failure as it coincides with the final step of the dielectric degradation. Although Weibull and log-normal laws were both used in the past, the pertinence of the Weibull distribution is no longer discussed [63, 67].

Considering a Weibull variable t, the probability density function is:

$$f(t;\beta,\eta) = \frac{\beta}{\eta}\cdot\left(\frac{t}{\eta}\right)^{\beta-1}\cdot\exp\left(-\left(\frac{t}{\eta}\right)^{\beta}\right) \tag{3.8}$$

In equation (3.8), β is the so-called weibull slope or the shape parameter and η is the characteristic lifetime at 63% or the scale parameter.
Regarding the gate-oxide breakdown physics, $f(t)$ defines the probability density of a breakdown event occurrence at a given time t after onset of breakdown conditions. Hence, it is worth to define the percentile of failing devices in relation to the electrical stress. The integration of $f(t)$ defined in (3.8) on the interval $[0, t]$ yields:

$$F(t;\beta,\eta) = \int_0^t f(t)dt = 1 - \exp\left(-\left(\frac{t}{\eta}\right)^{\beta}\right) \tag{3.9}$$

The statistic of the gate-oxide breakdown is described by equation (3.9). $F(t)$ is the cumulative failure probability. In other words, it evaluates the population fraction of failure at age t.

The Weibull scale is particularly convenient to plot a Weibull distribution and is defined as follows:

$$W(F) = \ln(-\ln(1-F)) \tag{3.10}$$

The combination of equation (3.9) and equation (3.10) yields:

$$W(F) = \ln\left[-\ln\left[\exp\left(-\frac{t}{\eta}\right)^{\beta}\right]\right] = \beta\ln(t) - \beta\ln(\eta) \tag{3.11}$$

Plotting $W(F)$ against $ln(t)$ yields a straight line with slope β and y-intercept $-\beta \cdot ln(\eta)$. Arbitrary Weibull slopes are illustrated in figure 3.5.

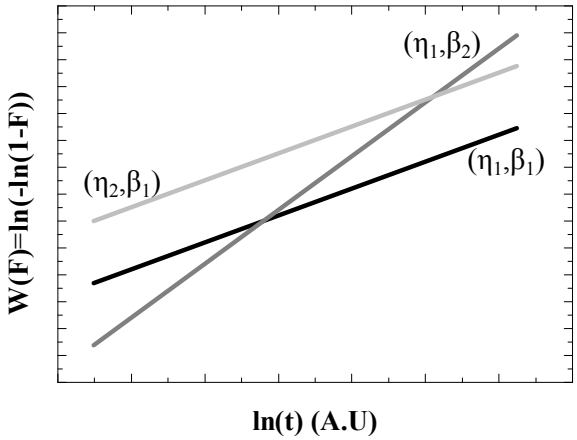


Figure 3.5: Weibull slopes with different β and η.

The impact of β and η are clearly emphasized. As mentionned previously, it is commonplace to characterize T_{BD} values using a Weibull statistic. In this case, the variable t is defined as T_{BD}. A set of capacitors are characterized using a constant voltage stress as shown in figure 3.6. Of course, the higher the number of devices, the most reliable the statistical distribution. A set of 50 devices can be assumed consistent.

3.2.2.2 Practical statistical study on antifuse bitcells

The reliability of the gate-oxide can be evaluated by means of characterizing the degradation under an electrical stress. This is the purpose of Time-Dependent Dielectric Breakdown (TDDB) experiments. A capacitor is stressed under a constant voltage until breakdown. A similarity with the programming operation of an antifuse bitcell can be noted. Typical TDDB measurements performed on a Drift antifuse bitcell (see in chapter 2, section 2.3.2.1) designed in the course of the project and fabricated in a logic 45-nm CMOS process are shown in figure 3.6.

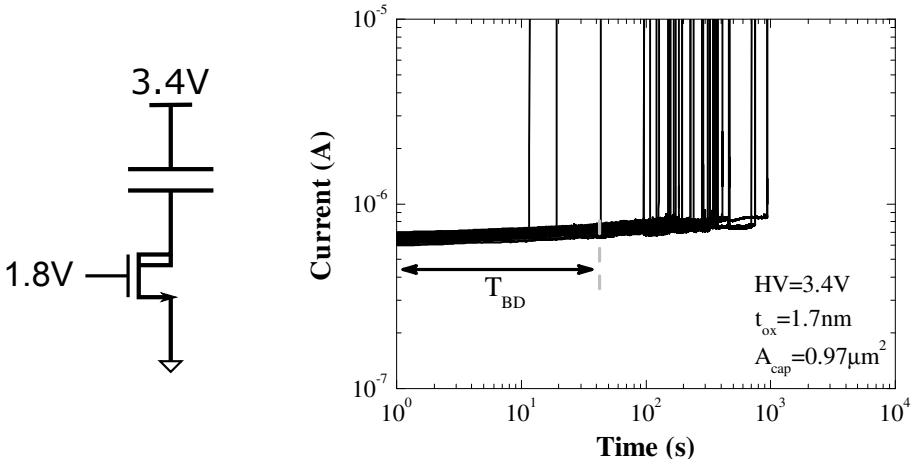


Figure 3.6: TDDB experiments on 45-nm drift antifuse bitcell. Experimental voltage conditions (a). Programming current measurements during a constant voltage stress (b).

72 bitcells were tested in this practical example. The programming voltage amplitude was chosen in the suitable range for a failure acceleration analysis (standard reliability test). When the DC voltage stress is applied across the device under stress, a wearout current of about 700nA flows through the dielectric. Then, the current rises up and the insulating property of the dielectric is irremediably lost. The time-to-breakdown is dispersed between 10^1s and 10^3s. Before commenting the statistical dispersion, it is worth to note that such a long programming time is obviously unacceptable for a memory application. This is the reason why the programming voltage of antifuse memory is much higher. However, the phenomenology observed with antifuse bitcells is the same as with regular MOS capacitors. Even though the stress voltage amplitude and the time-to-breakdown do not match the requirements for antifuse memory, studying the different approaches used in reliability is valuable. In a second step, the application of these studies and modeling approaches to antifuse bitcells are verified in chapter 4.

The Weibull distribution corresponding to the T_{BD} measurements performed on antifuse bitcells previously shown in figure 3.6 is plotted in figure 3.7.

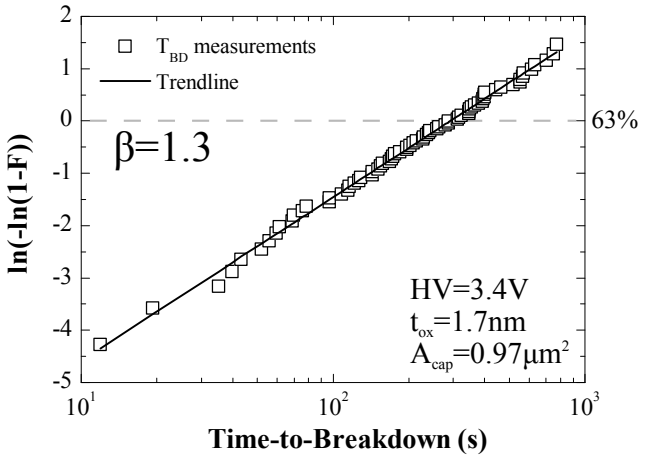


Figure 3.7: Time-to-Breakdown distribution for 72 devices (45-nm CMOS drift antifuse bitcell) programmed under 3.4V.

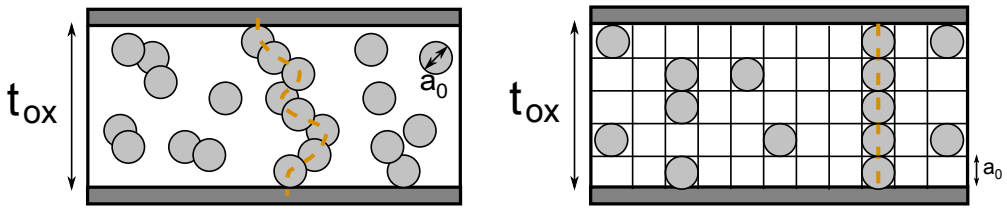
As expected, plotting $W(F)$ against $log(T_{BD})$ yields a straight line correctly fitted with a logarithmic trendline. The value of the Weibull slope is discussed in the next section. However, the wide dispersion of T_{BD} between roughly 10^1s and 10^3 s can be noticed.

T_{BD} values are here very large. When the programming conditions are studied, T_{BD} values are significantly smaller and the measurement setup needs to be adapted as detailed in section 3.4.

3.2.2.3 Percolation model

Even though the pertinence of the Weibull statistic is well accepted, a model taking into account the different parameters of a capacitor is needed.

Assuming that the gate-oxide breakdown is a local loss of insulating properties, it can be approached by a conduction path between the electrodes of a capacitor [68]. The so-called *percolation model* relies on the random generation of spherical defects within the dielectric material stressed under a constant voltage stress [54]. A conduction between two elements is achieved if one overlaps onto another. Once the electrodes are connected by a chain of defects, the breakdown state is obtained. This chain of defects is also called the *percolation path*. An illustration of the original percolation model is depicted in figure 3.8(a). Since 1995, the statistical model of T_{BD} is still studied. An example of another approach is shown in figure 3.8(b).



(a) (b)

Figure 3.8: The original percolation model [54]. A modification of the percolation model [69].

The definition of the percolation model involves the diameter of a spherical defect a_0 (see in figure 3.8(a)) or the size of a cell (see in figure 3.8(b)). The breakdown state is defined by a critical breakdown defect density per unit of area N_{BD}. An equation of N_{BD} is [69]:

$$N_{BD} = \frac{t_{ox}}{a_0^3} \cdot \exp\left[-\frac{1}{\beta \cdot t_{ox}} \ln\left(\frac{A_{ox}}{a_0^2}\right)\right] \tag{3.12}$$

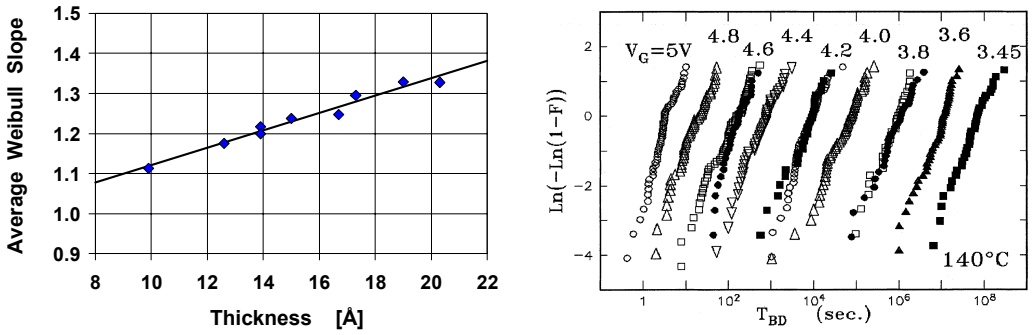
N_{BD} depends on the defect size a_0, the gate-oxide thickness t_{ox}, the Weibull slope β and the capacitor area A_{ox}. The expression of β is [69]:

$$\beta = \alpha \frac{t_{ox}}{a_0} \tag{3.13}$$

Several values of a_0 and α are reported, for instance, 0.9nm [55], 1.83nm [70], 0.4nm [71] for a_0 and 0.5 [55], 1 [63] for α. The parameter α is identified from measurements.

Many satisfying models based on the latter approach are reported in literature. Although other approaches are possible, many authors converge on the pertinence of the Weibull statistic. The distribution depends on the device area and on the Weibull slope which is related to t_{ox}. Moreover, the Weibull distribution is voltage-independent. The practical modeling of the Weibull slope according to the gate-oxide thickness is a significant result from the percolation approach.

The influence of t_{ox} on β is shown in figure 3.9(a). Distributions of T_{BD} are plotted in figure 3.9(b) for different voltage stresses. The Weibull slope is constant as defined in equation (3.13), .



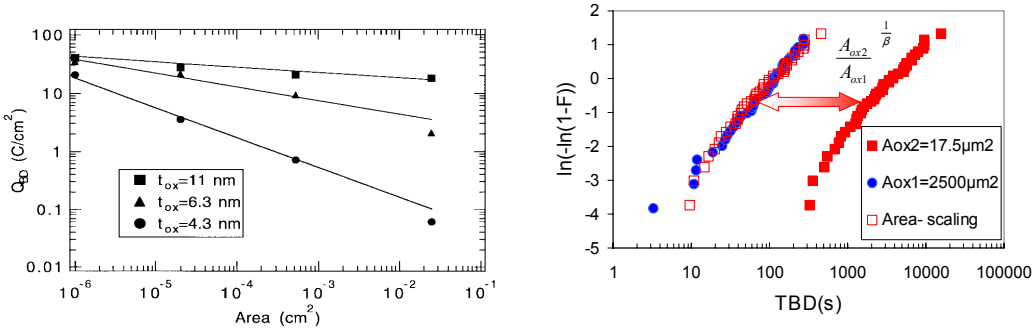
(a) (b)

Figure 3.9: Modeling and measurements of the Weibull slopes for various oxide thicknesses [71] (a). T_{BD} distributions for various stress voltages [72] (b).

The modeling of T_{BD} according to the device area is also a major property of the percolation model. The capacitor area has a direct impact on the charges injected before the failure. The *charge-to-breakdown* Q_{BD} is defined as the integration of the current density on the interval $[0, T_{BD}]$:

$$Q_{BD} = \int_0^{T_{BD}} J_g(t)\, dt \tag{3.14}$$

Q_{BD} is plotted as a function of A_{ox} in figure 3.10(a). A large capacitor exhibits a smaller Q_{BD}, reflecting a shorter T_{BD}. This observation is confirmed in figure 3.10(b). T_{BD} distributions of capacitors featuring very different area are compared.



(a) (b)

Figure 3.10: Q_{BD} plotted versus A_{ox} [55] (a). T_{BD} distributions of two capacitors [73] (b).

The large area capacitor exhibits a shorter T_{BD} than the small area capacitor. In fact, the influence of the capacitor area is associated to the probability of creating a percolation path and to the random distribution of defects. In other words, the

critical defect density is reached in a shorter time for a large device than for a short device due to a higher defect generation rate.
The percolation model allows the normalization of T_{BD} distributions according to the device area. This is the definition of the so-called *area-scaling law*.

$$\frac{T_{BD1}}{T_{BD2}} \approx \frac{Q_{BD1}}{Q_{BD2}} = \left(\frac{A_{ox2}}{A_{ox1}}\right)^{\frac{1}{\beta}} \tag{3.15}$$

The area scaling demonstrates that the percolation model is 3-D. In other words, defects are randomly generated in the dielectric along t_{ox} and across A_{ox}, i.e. in volume. A distribution normalized using the area-scaling law is plotted in figure 3.10(b). T_{BD} can be then extrapolated to different capacitor areas.
The percolation model illustrates the degradation process involved in the gate-oxide breakdown. Although the impact of T_{ox} and the capacitor area are modeled, it is seen that the breakdown is finally a local phenomenon.

3.2.3 Voltage acceleration of time-to-breakdown

The percolation model presented in section 3.2.2.3 is useful to predict T_{BD} according to t_{ox} and A_{ox} and to demonstrate the pertinence of Weibull distributions. However, it was shown that the critical breakdown defect density is voltage-independent whereas T_{BD} is accelerated by the stress voltage.
The modeling of the voltage-acceleration of T_{BD} is one of the most discussed and reported topic in reliability. It is a major concern for the case of antifuse bitcell programming because of the very high voltage range.

3.2.3.1 Empirical approach

The degradation of devices is accelerated by applying a voltage much higher than the nominal condition. Then, reliability results are projected in a nominal voltage range and T_{BD} of years are extrapolated. The gate-oxide breakdown is a "weakest-link" type of failure. In other words, the failure of a single transistor leads to the failure of the whole chip. The relevancy of the voltage-acceleration law of T_{BD} is therefore crucial to guarantee the reliability of electronics products for years. Examples of projection using several voltage-acceleration laws are plotted in figure 3.11.

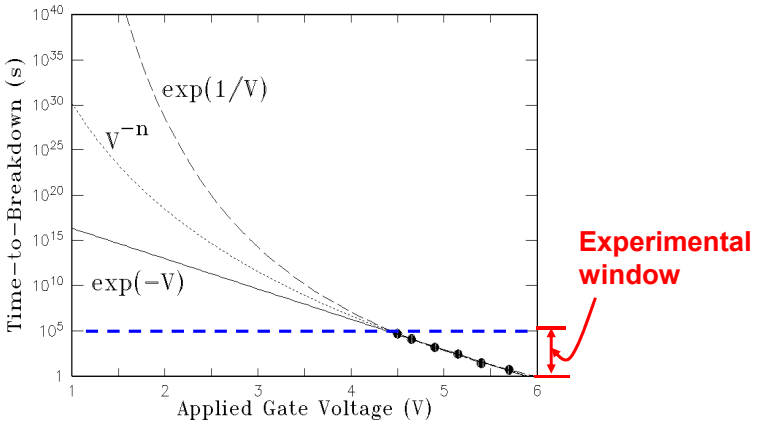


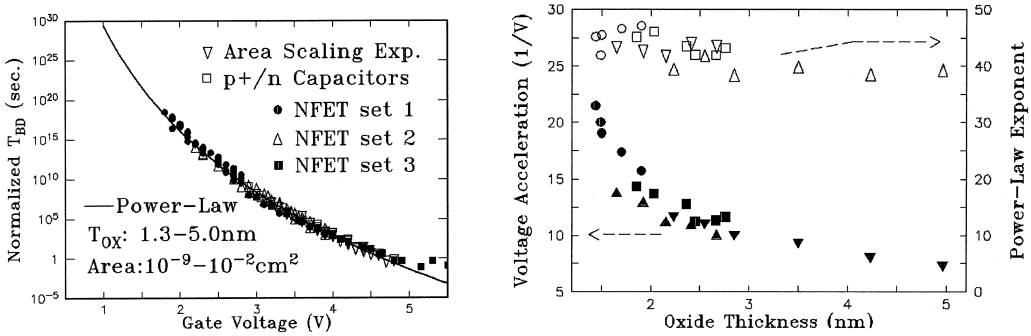
Figure 3.11: Lifetime projection using different laws [74]

The impact of the voltage-acceleration law is clearly emphasized in this example. Measurements are performed in an experimental time window. T_{BD} should be in a range between an acceptable test time and the bandwidth limit of the test equipment. In this example, the experimental time window is between 1s and 10^5s. Then, the measurements are fitted with mathematical laws such as exponential or power. These laws are used to project the lifetime in nominal voltage conditions. It can be clearly noticed in figure 3.11 that an irrelevant law leads to severe discrepancies whereas the trendline is correctly fitted against measurements in the experimental window.

An empirical approach consists in extending the experimental window in order to perform a more accurate curve fitting. For thick oxide ($t_{ox} > 6$nm), the high voltage stress led to T_{BD} projections using exponential laws. With the reduction of the gate-oxide thickness, the nominal and stress voltages were shifted towards lower amplitudes. As a consequence, it turns out that an extrapolation of T_{BD} using an exponential law exhibited discrepancies [72].

An interesting experimental evidence was pointed out by Wu *et al* in 2002 [72]. A voltage dependence of T_{BD} and Q_{BD} was demonstrated after testing hundreds of devices. Authors demonstrated that modeling the gate-oxide breakdown using an exponential-law violated the area-scaling property of the percolation model whereas this result remained valid using a power-law.

For $t_{ox} < 50\,\mathring{A}$, the voltage-acceleration of T_{BD} follows a power-law relation as plotted in figure 3.12(a). A typical power-law exponent of about -43 ± 3 has been experimentally determined for a range of t_{ox} from 1.5nm to 5nm. Only the voltage-acceleration factor varies. Both parameters are plotted in figure 3.12(b).



(a) Typical T_{BD} data versus the gate voltage for a variety of structures over a wide range of thickness and area [72].

(b) Voltage-acceleration factors and power-law exponents as a function of t_{ox} [70].

Figure 3.12: Illustration of the power-law voltage acceleration of T_{BD}.

The experimental demonstration of the relevancy of the power law as a model of T_{BD} made a connection between the percolation model and experimental observations. An empirical definition of T_{BD} using a power law can be derived as follows:

$$T_{BD} = \alpha V_g^{-n} \tag{3.16}$$

There is an interest to discuss a physics-based identification of parameters α and n in contrary to an empirical approach as in [70, 72, 75]. The aim of physics-based models is to achieve a more accurate lifetime extrapolation.

3.2.4 Physics-based models

Relevant trends on a voltage-acceleration laws of T_{BD} are identified using an empirical approaches. However, the narrow experimental window led to unacceptable errors on lifetime extrapolation. In this respect, having a trustworthy lifetime model for MOS devices is valuable. Strong groundworks have been reported such as the percolation model and the empirical power law for ultrathin oxide. However, physical origins of the phenomena observed in experiments raised many interrogations. Sustained efforts are made towards the modeling of microscopic breakdown mechanisms in order to provide accurate projection laws. The main models reported in literature are presented in this section.

3.2.4.1 Anode Hole Injection: 1/E model

A model based on hot hole injection from the anode (AHI) was introduced by Schuegraf *et al* in 1994 [76]. Authors assumed that incident energized electrons tunneling through the gate-oxide are able to transfer their energy at the anode edge

to holes. Then, the hot holes tunnel back into the dielectric. The degradation mechanism is depicted in figure 3.13.

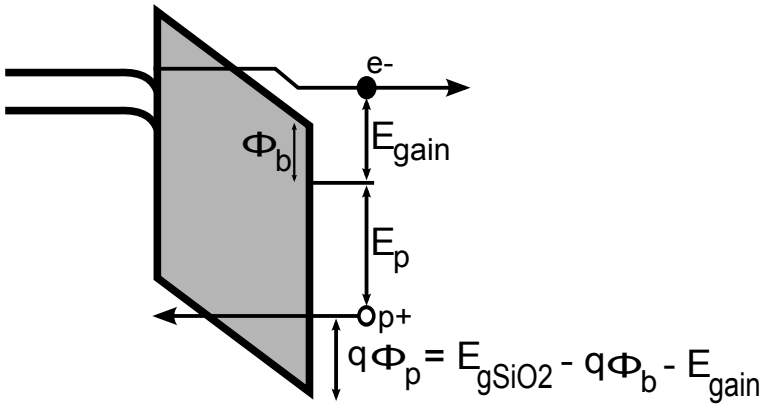
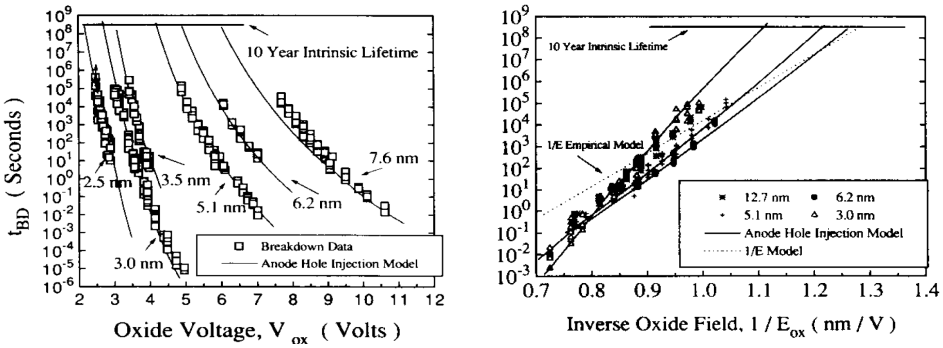


Figure 3.13: Diagram of the Anode Hole Injection process [76]

$E_{gSiO2} = 9eV$ is the forbidden gap of SiO_2, E_{gain} is the energy released at the anode edge by tunneling electrons, $q \cdot \phi b = 3.2eV$ is the electron injection barrier height and $q \cdot \phi p$ is the hole injection barrier height. The equation of $q \cdot \phi p = 4.6eV$ is valid only for Fowler-Nordheim current therefore $V_{ox} > \phi b$. Otherwise, the energy of electrons is merely $E_{gain} = V_{ox}$ for $V_{ox} < \phi b$ in a direct tunneling regime.
Assuming a Fowler-Nordheim conduction regime, an expression of T_{BD} can be derived as follows:

$$T_{BD} \propto exp\left(\frac{G}{E_{ox}}\right) \tag{3.17}$$

The expression of T_{BD} derived in equation (3.17) exhibits a dependence in $\frac{1}{E_{ox}}$. This is the reason why the AHI model is also commonly called "$\frac{1}{E}$" model.
The AHI model was built for thick oxide from 2.5 to 13nm and for a Fowler-Nordheim conduction regime. Authors justified this restrictions by the fact that observation of direct tunneling current is limited to oxides thinner than 5nm due to the experimental window.
Lifetime projections using AHI and fitted against measurements are plotted in figure 3.14.



(a) (b)

Figure 3.14: Voltage dependence of breakdown lifetime predicted using AHI model (a). Inverse field dependence of breakdown lifetime compared to an empirical 1/E model (b) [76].

Due to the ultrathin oxide used in state-of-the-art MOS technologies ($t_{ox} < 2nm$), the AHI model does not seem applicable.

3.2.4.2 Thermo-Chemical: E model

This model was proposed in the 80s by McPherson and reviewed in the 90s [77]. The so-called thermo-chemical, field-enhanced bond breakage or "E-model" suggests that the defect generation in a dielectric is a field-driven process. Basically, the interaction of the electric field with weak atomic bonds is considered in this model as the main cause of gate-oxide breakdown.
According to the E-model, T_{BD} depends on the electrical field as defined in the following equation:

$$T_{BD} \propto exp\,(\gamma E_{ox}) \tag{3.18}$$

In equation (3.18), γ is an acceleration factor. The field-driven process is emphasized by the E_{ox} term in the exponential, this is also the reason why the thermo-chemical model is also called "E-model".
A comparison between the "E" and the "$\frac{1}{E}$" model is shown in figure 3.15. Both models are fitted against experimental data.

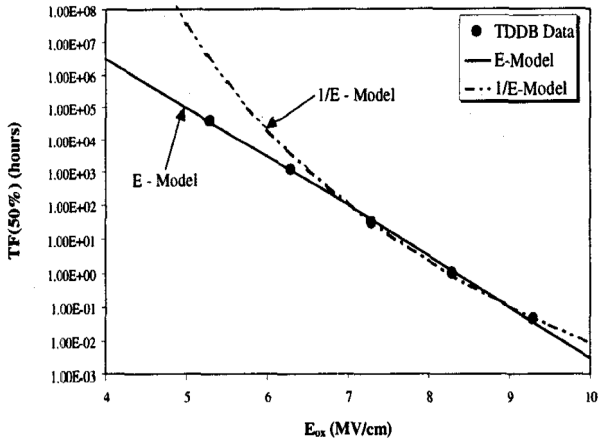


Figure 3.15: Comparison of E-model and 1/E-model fitted to TDDB data [78].

Even though a better fit is obtained in this example using the E-model, a trend towards a power-law voltage-acceleration of T_{BD} instead of an exponential law was reported in numerous works on ultrathin oxides. Consequently the validity of the E-model can be questioned. To conclude, it can be noted that authors recently reviewed the E-model in 2007 in order to explain a possible deviation of TDDB data from a pure exponential law [79].

3.2.4.3 Hydrogen release

The hydrogen release mechanism was reported by DiMaria *et al* in 1995 [80]. The underlying principle is based on the release of hydrogen species at the interface of the dielectric. An illustration of the mechanism is depicted in figure 3.16.

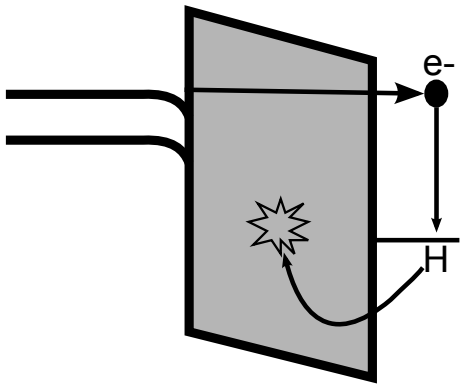


Figure 3.16: Diagram of the hydrogen release process.

Hydrogen species are located at the interface between the silicon substrate and the silicon dioxide due to an hydrogen annealing step during manufacturing. This

process step is used to passivate Si dangling bonds by forming $Si-H$ bonds at the interface. However, these bonds are weak.
The gate-oxide degradation process by hydrogen release can be divided in three main steps:

1. **Hydrogen release:** due to the energy given by electrons tunneling through the gate-oxide whether in direct of Fowler-Nordheim conduction mode, $Si-H$ bonds are broken and hydrogen atoms are released at the gate/dielectric interface, thereby creating an interface state.

2. **Hydrogen diffusion:** after the release, hydrogen species are diffused in the dielectric layer.

3. **Hydrogen recombination:** the hydrogen atoms diffused into the SiO_2 layer can be combined with oxygen vacancies yielding defects in the oxide until the creation of a conduction path.

Wu *et al* mentioned a possible link between the T_{BD} power-law dependence and an underlying hydrogen release mechanism, especially about the origin of the large exponent [72]. However, such a result could not be claimed due to a lack of experimental data.
Ribes *et al* have demonstrated the relevancy of gate-oxide degradation due to hydrogen release for ultrathin oxides [73]. The author built up the so-called Multi Vibrational Hydrogen Release (MVHR) model in which the physical mechanism involved in the gate-oxide breakdown is the release of hydrogen species either at Si/SiO_2 or the Si/polysilicon interface (gate or substrate carrier injection) by a multi-vibrational excitation. In other words, several carriers contribute to the breakage of a single $Si-H$ bond.
The expression of T_{BD} and Q_{BD} using a power law is a major outcome. Indeed, the origin of the power-law voltage acceleration of the gate-oxide breakdown can be connected to the $Si-H$ desorption energy E_{th}.

$$T_{BD} \propto V_g^{\beta} \qquad Q_{BD} \propto V_g^{\delta} \tag{3.19}$$

$$\beta = n_{fin} \cdot \frac{E_{th}}{n \cdot \sigma} + n_{Ig} \cdot \frac{E_{th}}{n \cdot \sigma} \qquad \delta = n_{fin} \cdot \frac{E_{th}}{n \cdot \sigma} + n_{Ig} \cdot \left(\frac{E_{th}}{n \cdot \sigma} - 1 \right) \tag{3.20}$$

n_{fin} is the power law exponent identified from the energy dependence of the inelastic fraction of tunneling electrons [81], E_{th} is the $Si-H$ desorption energy, σ is the energy of stretching exited states mode and n is the number of energy quanta transferred by the vibrational excitation. The wearout current is also taken into account and modeled using a power-law exponent n_{Ig} extracted from prior measurements.

Without detailing in depth every parameter, a link between the power-law exponent and the $Si-H$ bond physics can be noticed in equations (3.19). Models fitted against measurements are shown in figure 3.17 for NMOS and PMOS capacitors either in accumulation or inversion regime.

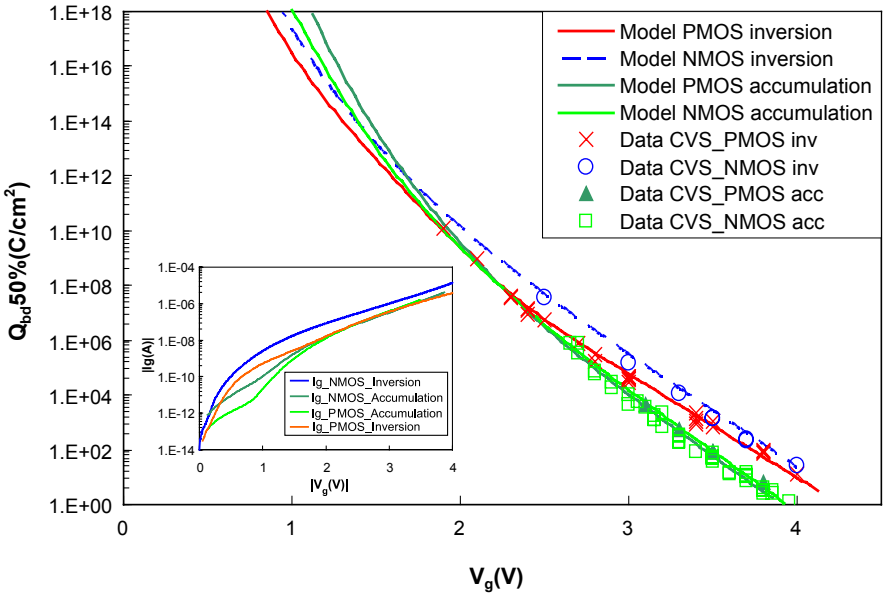


Figure 3.17: Charge-to-breakdown versus V_g for all stress configuration [82].

The MVHR model brought a mathematical law to extrapolate the lifetime for ultrathin oxides (corresponding to the 45-nm technology and beyond). The model is applicable to the reliability of PMOS and NMOS transistors either in accumulation or inversion regime.
Finally, this model provides accurate lifetime extrapolation over a wide T_{BD} range of validity with respect to reliability.

3.2.5 Conclusion and perspective for antifuse bitcells

The gate-oxide breakdown is a major concern regarding reliability issues of CMOS technologies. Research efforts have been made in order to understand and to model this physical phenomenon. As presented in this section, the statistical approach by the percolation model is a strong groundwork.
For now, research works are focused on a deterministic approach in order to point out the underlying physics involved in dielectric breakdown from a reliability point of view. Three models were presented in section 3.2.4.
The anode hole injection and hydrogen release models are based on the interaction between carriers and the gate-oxide. As explained in 3.2.4.1, the AHI model is proven valid for $2.5\text{nm} < t_{ox} < 13\text{nm}$. Due to the ultrathin oxide thickness used in

state-of-the-art CMOS technologies ($t_{ox} < 2nm$), this model does not seem suitable for an accurate time-to-breakdown (T_{BD}) modeling.

The exponential field-dependence of T_{BD} is modeled by the thermo-chemical model as an action of the electric field on atomic bonds within the gate-oxide. The role of the carriers is considered as secondary. Moreover, the T_{BD} voltage-acceleration defined as an exponential is in contradiction with the power-law dependence observed experimentally for ultrathin oxide. In addition, inconsistencies were observed with the percolation model as the violation of the area scaling law. However, it can be noticed in figure 3.11 that the E-model is the most conservative.

Models based on hydrogen release are a possibility to justify the empirical power law voltage-acceleration of T_{BD} even for low voltages and ultrathin oxides. The Multi Vibrational Hydrogen Release (MVHR) model appears as a very promising candidate even for high-K dielectrics [83]. However, the voltage range in which the model is fitted against measurements is limited between 2V and 4V in figure 3.17.

The present work aims at characterizing the gate-oxide breakdown occurring in antifuse bitcells programmed under a high voltage in order to point out possible optimization. Since the validity of physics-based Time-Dependent Dielectric Breakdown (TDDB) models has not been proven beforehand for such voltage conditions, the accurate validation would take a considerable time. Consequently, efforts will be made towards the description of the gate-oxide breakdown and the identification of model parameters using a trend line, e.g. a power law. The first task is to set up experimental methodologies to perform fast T_{BD} characterizations under high voltages. Thus, data can be fitted against models in order to identify a pertinent law. Characterization methodologies are discussed in section 3.4.

3.3 Antifuse bitcells and high-K dielectrics

For more than ten years, the high performance of MOS devices was enhanced by reducing the gate length and the the gate-oxide thickness. The transistors were entirely made of silicon as a SiO_2 dielectric and a polysilicon gate. However, the ultrathin gate-oxide leads to a penalty in leakage current, i.e. in power consumption. This is the reason why decreasing the dielectric thickness is no longer suitable for the next CMOS technology nodes, e.g. 32-nm, 28nm, 20-nm.

A solution to overcome the limitation of fully-silicon devices is to use a material featuring a higher permittivity than SiO_2. Such dielectrics are also known as High-K. In fact, the performance of a MOS transistor depends on the gate-oxide capacitance C_{ox} as defined in equation (3.21).

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \tag{3.21}$$

Where ε_{ox} is the permittivity of the dielectric and t_{ox} the dielectric thickness. The reduction of t_{ox} is therefore justified. However, C_{ox} can be also by increased by means of replacing the SiO_2 dielectric by a high-K material. In addition, the polysilicon gate can be replaced by a metal gate in order to avoid a deep depletion in inversion regime.

The main challenge is the complexity of the manufacturing process. The advantage of fully-silicon devices was the quality of the interfaces between the SiO_2 layer, the substrate and the polysilicon gate.

A schematic illustrating the integration of a fully-silicon and a high-k capacitor is shown in figure 3.18.

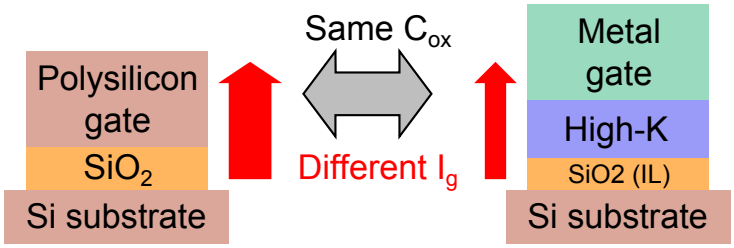


Figure 3.18: Integration of a fully-silicon and high-k capacitors and impact on the leakage current [84].

A SiO_2 interface layer is used between the substrate and the high-K dielectric in order to obtain a better interface quality and therefore a better control of the threshold voltage.

The benefits of using a high-K dielectric can be emphasized by calculating the Equivalent Oxide Thickness (EOT) of the interface and high-K layer as in equation (3.22).

$$EOT = \frac{\varepsilon_{SiO2}}{\varepsilon_{SiO2}} t_{IL} + \frac{\varepsilon_{SiO2}}{\varepsilon_{HK}} t_{HK} \tag{3.22}$$

A possible high-K material is the hafnium dioxide HfO_2. Its dielectric constant is 25 whereas the dielectric constant of SiO_2 is 3.9. Assuming a SiO_2 interface layer of 1nm and a HfO_2 layer of 4nm, the EOT is:

$$EOT = \frac{3.9}{3.9} \cdot 1nm + \frac{3.9}{25} \cdot 4nm = 1.62nm \tag{3.23}$$

The performance of the dielectric formed with an interface layer of 1nm plus a high-K layer of 4nm is equivalent to a single SiO_2 layer featuring a physical thickness of

1.62nm. The major advantage is the reduction in leakage current as shown in figure 3.18.
The reduction in gate leakage current using a high-k dielectric is emphasized in figure 3.19.

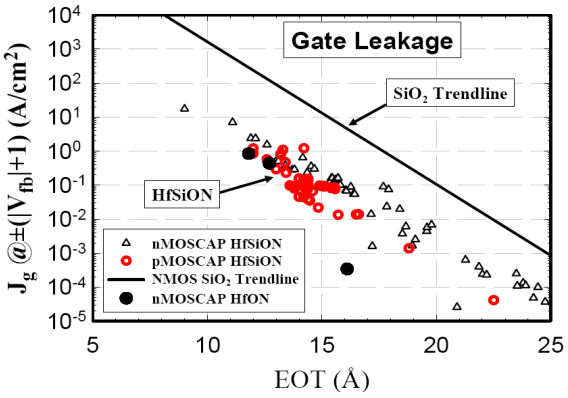
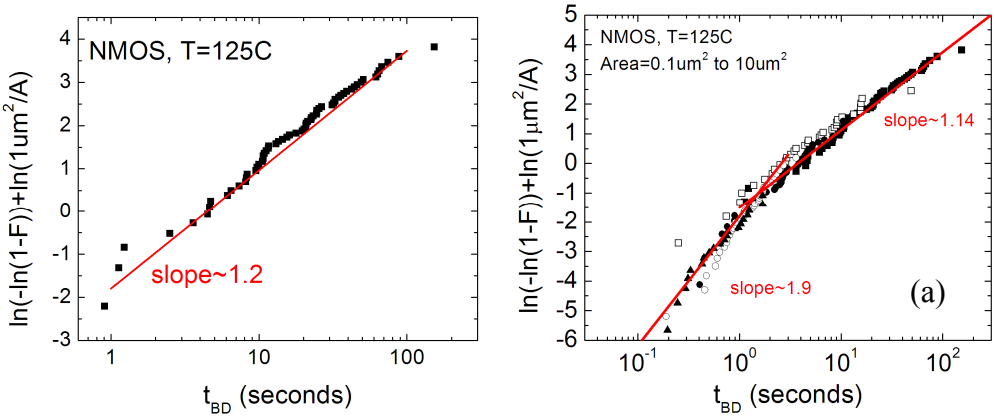


Figure 3.19: Reduction of leakage current using HfSiOn and HfON gate dielectric [85]

The leakage current of different MOS capacitors featuring hafnium silicon oxynitride HfSiON and hafnium oxynitride HfON is compared with a SiO_2 trendline. It can be clearly seen that the leakage current is one order of magnitude lower using a high-K dielectric. Hence, the question about a better robustness of the dielectric in antifuse bitcell applications can be raised.

3.3.1 High-K dielectric breakdown

The reliability of high-K dielectric is characterized to ensure the functionality of future ICs. Since an antifuse memory is fully compatible with a standard process, the antifuse capacitor features a high-K dielectric, a SiO_2 interface layer and a metal gate. Consequently, there is an interest in understanding the breakdown physics of this complex stack.
The methodology presented in this chapter can be used as a first approach for providing insights into the underlying physics. This is the reason why reliability engineers performed TDDB characterizations on high-K/metal gate devices. Time-to-Breakdown distributions are plotted in figure 3.20.



(a) (b)

Figure 3.20: Experimental TDDB distributions for (a) $0.1\mu m^2$ and for (b) different areas scaled to a $1\mu m^2$ reference area [86].

The distribution shown in figure 3.20(a) is consistent with a Weibull distribution. Besides the slope corresponds to the value found for ultrathin SiO_2 gate-oxide.
As explained in section 3.2.2, the area scaling is a key property of the percolation model. The defect generation is random in the gate-oxide and across the area. In other words, distributions for different areas can be scaled to a reference. In the example shown in figure 3.20(b), the Weibull slopes is not constant for an area range from $0.1\mu m^2$ to $10\mu m^2$. Wide capacitors exhibit a slope of 1.9 whereas small capacitors exhibit a slope of 1.14.
The cause of the bimodal Weibull distribution is, according to authors, the result of a faster defect generation in the high-K layer [86]. The breakdown is dictated by the interface layer for small capacitors. This is the reason why the Weibull slope is consistent with the value characterized for ultrathin oxide whereas the higher degradation rate in the high-K layer is more significant for large capacitors.
The complexity in TDDB model of such dielectric stacks is emphasized in this example. The lifetime projection is different according to the area of the device.
The modeling of the time-to-breakdown according to the voltage stress is also discussed in reliability. Like for fully-silicon devices, the aim of a voltage-acceleration model is to determine a T_{BD} value according to a stress voltage using a mathematical law, e.g. power, exponential. An example is shown in figure 3.21.

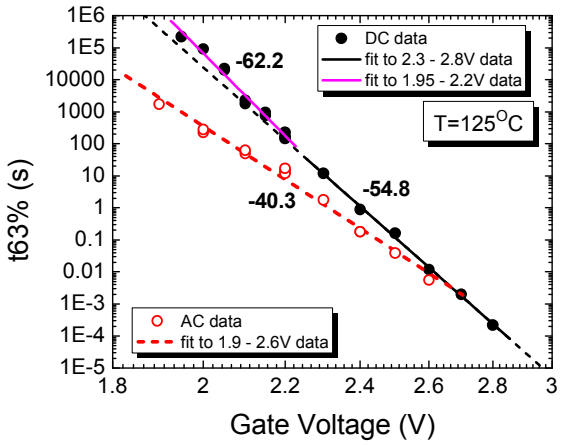


Figure 3.21: Experimental T_{BD} measurements fitted to a power-law dependence. For AC stress (100Hz, $V_{recovery} = -2V$) a significantly lower acceleration factor (-40.3) is obtained [87].

Considering the DC stress conditions, it can be seen that the T_{BD} values cannot be fitted by a single power law model. An exponent of n=-54.8 is identified for $V_g > 2.2V$ whereas n=62 for $V_g < 2.2V$. Furthermore, trends on T_{BD} values is completely different in AC stress. The acceleration factor is n=-40.3 and keeps the same value for V_g from 1.95V to 2.6V.

The difference observed on the voltage-acceleration of T_{BD} for different stress voltage range or stress mode is due to a phenomenon of charge-trapping in the high-K dielectric. Indeed, authors in [87] pointed out significant shift of the threshold voltage. Even though the power-law dependence of T_{BD} is relevant, an accurate model should take into account the charge-trapping mechanism according to the stress voltage amplitude and mode.

3.3.2 Perspectives

As mentioned previously, an antifuse bitcell should be fully compatible with the CMOS logic process. As a consequence, a 32-nm antifuse capacitor features a high-K/metal gate stack.

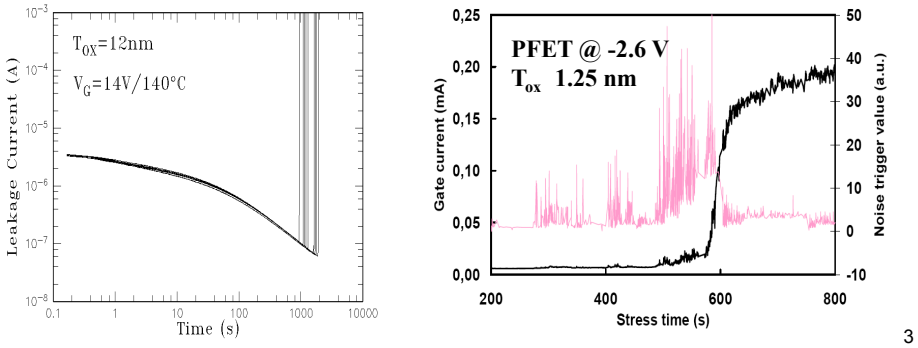
The breakdown physics of ultrathin SiO_2 layers is well-known in reliability. However, the accurate understanding and characterization under the high programming voltage used in antifuse memory is rarely reported. Consequently, it would be difficult to model the TDDB of high-k / metal gate stacks stressed under high voltage. Like for SiO_2, efforts will be focused on description of the breakdown mechanisms.

3.4 Time-to-Breakdown characterization

The high voltage and short time-to-breakdown required to meet acceptable performance of antifuse memories are out of the range of conventional Time-Dependent Dielectric Breakdown characterizations. According to the reported programming time presented in chapter 2, a measurement range corresponding to T_{BD} down to at least $1\mu s$ is suitable. In addition to TDDB experiments, gate current transient measurements are also necessary to characterize the degradation of the gate-oxide during the wearout phase.
After a literature review about the gate-oxide breakdown event, different experimental setups and characterization results are presented in this section. Furthermore, the different methodologies are tested on antifuse bitcells designed and fabricated in the course of this project.

3.4.1 Gate-oxide breakdown event

With the continuous downscaling of MOS devices and the reduction of the gate-oxide thickness, different evolutions of gate-current appeared in TDDB experiments.
The gate current is shown in figure 3.22 for two MOS capacitors with respectively $t_{ox1} = 12nm$ and $t_{ox2} = 1.25nm$.



(a) (b)

Figure 3.22: Transient gate current measurements along a constant voltage stress of two MOS capacitors with respectively $t_{ox1} = 12nm$ and $t_{ox2} = 1.25nm$ [74].

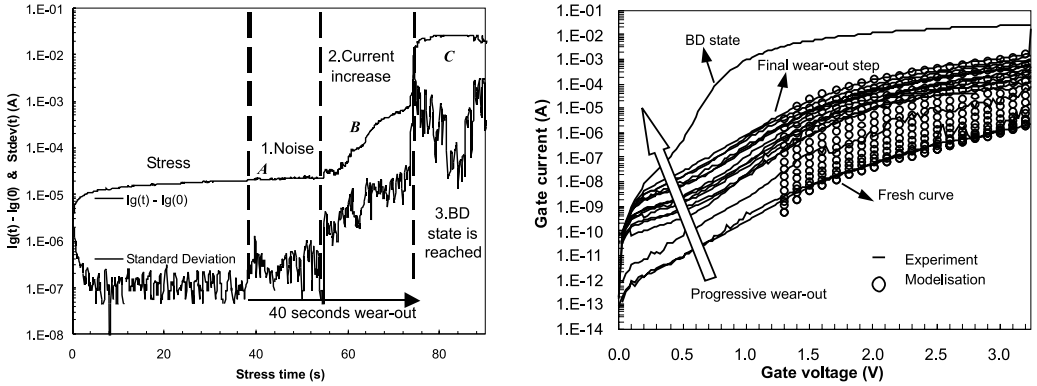
The gate current shape measured on the two devices is clearly different. The stress voltage was adapted in order to remain in a suitable experimental window, i.e. T_{BD} in a measurable range. In figure 3.22(a), a hard breakdown occurs after a wearout current decay due to charge trapping. In figure 3.22(b), the wearout current becomes noisier before the breakdown event. Then, the failure arises during a third phase in which the gate-current increases somewhat continuously and noisily.

With the continuous down scaling of the gate-oxide thickness, it turned out that the definition of the gate-oxide breakdown by a steep current growth is no longer valid. Indeed, the transition between an insulating and a conducting state is more progressive. As a consequence, the breakdown hardness has been studied intensively for the last ten years on various technology nodes.

3.4.1.1 Breakdown modes

Since the breakdown electrical signature has changed, the different state observed during a constant voltage stress must be clearly defined. As a consequence, two breakdown modes can be identified [88].

- **Hard breakdown (HBD):** the gate-current growth reaches the compliance of the test equipment (see in figure 3.22(a)).
- **Progressive breakdown (PBD):** the gate-current increases continuously and noisily, precursor state of a hard breakdown (see in figure 3.23(a)).



(a) (b)

Figure 3.23: Gate-current measurements for a 1.8-nm gate-oxide (Area $= 40000\mu m^2$, T = 125°C, 0.12μm CMOS technology) [89].

For gate-oxides thinner than 2nm, the observed breakdown mode is progressive.

3.4.1.2 Focus on soft and progressive breakdown modes

The progressive breakdown is a potential useful signature for antifuse due to its irreversible nature and a post-breakdown current lower than HBD. Hence, the noise annunciating a breakdown and the intermediate current level can be innovative criteria for a programmed bitcell.
Progressive or soft breakdown was reported for the first time by Lee *et al* [90]. They observed particular gate-current shapes on 4nm gate-oxide devices. Then,

many scientists worked on this phenomenon in order to understand the impact of this failue mode on the functionality of integrated circuits. An example of lifetime extension is shown in figure 3.24.

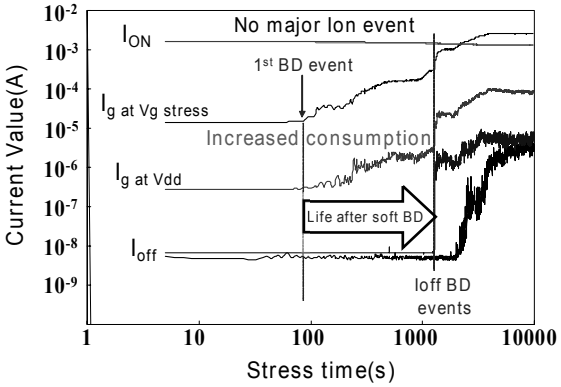


Figure 3.24: Typical post-breakdown transistor current degradation (PMOS $1 \times 0.6\mu$m, $T_{ox} = 1.3$nm) [91].

A first breakdown event occurs leading to an increase in power consumption. Nevertheless, the device remains operational. From a reliability point of view, the failure is obtained for a hard breakdown. Lifetime models presented in 3.2.4 are based on T_{BD} measurements. Looking at the different breakdown modes, it has been demonstrated that only the noise detection is consistent with the percolation model [88]. Hence, the model is built up according to progressive or soft T_{BD} measurements and can be extended towards hard breakdown failures [91, 92].

Different interpretations of the breakdown physics are reported in literature. One of the major interrogation was the correlation between soft or progressive and hard breakdown [93–97]. In fact, the emergence of progressive and soft breakdown correlated with the reduction of the gate-oxide thickness can be explained by studying the experimental time window (see in 3.2.3).

Considering Δt as the sampling time of test equipment, the case whether the breakdown is detected as hard or progressive is set by the delay duration between the noise appearance and the steep current increase. This approach is illustrated in figure 3.25.

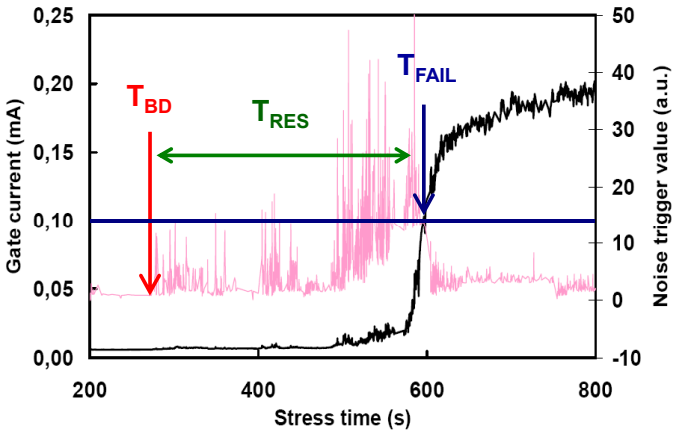


Figure 3.25: [74]

The delay duration (T_{res}) depends on the temperature and the voltage. Two cases can be identified:

- If $\mathbf{T_{res} > \Delta t}$, a progressive breakdown is detected.
- If $\mathbf{T_{res} < \Delta t}$, a hard breakdown is detected.

Furthermore, the emergence of progressive breakdown according to gate-oxide thickness and the stress voltage can be approached as depicted in figure 3.26.

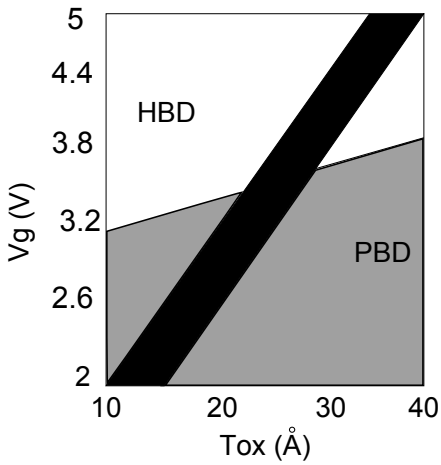


Figure 3.26: Grey area: progressive breakdown. White area: hard breakdown, Black area: stress voltage amplitude required to measure T_{BD} in an experimental window such that $10s < T_{BD} < 6000s$ [98].

The diagram emphasizes a trend in progressive breakdown emergence according to the downscaling of the gate-oxide thickness. Indeed for $t_{ox} < 20$ Å, only progressive breakdown are detected considering the experimental window defined by the author.

Like TDDB modeling, the breakdown modes are studied under low voltage stress. Therefore the transient evolution of antifuse bitcell programming current must be measured in order to classify the gate-oxide breakdown. Moreover, the increase in noise during a progressive breakdown event can be used as a signature for a low-power programming solution.

3.4.2 Antifuse bitcell characterization methods

Constant voltage stress experiments are suitable to perform TDDB characterizations on antifuse bitcells in a minimum time range of 1s. An example was given in 3.2 figure 3.6. However, the targeted programming time for antifuse memories is about 10μs.

Experimental setups featuring an appropriate bandwidth are needed to perform TDDB characterizations and programming current measurements of antifuse bitcells under a high voltage stress. Measurements are performed on test structures designed and fabricated in the course of this Ph.D. work using the different experimental methodologies presented along this section. Then, conclusions on the suitability and the feasibility for antifuse bitcell characterizations are drawn.

3.4.2.1 DC voltage ramp

Even though antifuse bitcells are programmed under a constant voltage stress, a DC voltage ramp is a useful experimental method to evaluate the current over a wide voltage range.

Measurements performed on 6 drift antifuse bitcells designed fabricated in a logic CMOS 45-nm process ($t_{ox} = 17$Å) are shown in figure 3.27.

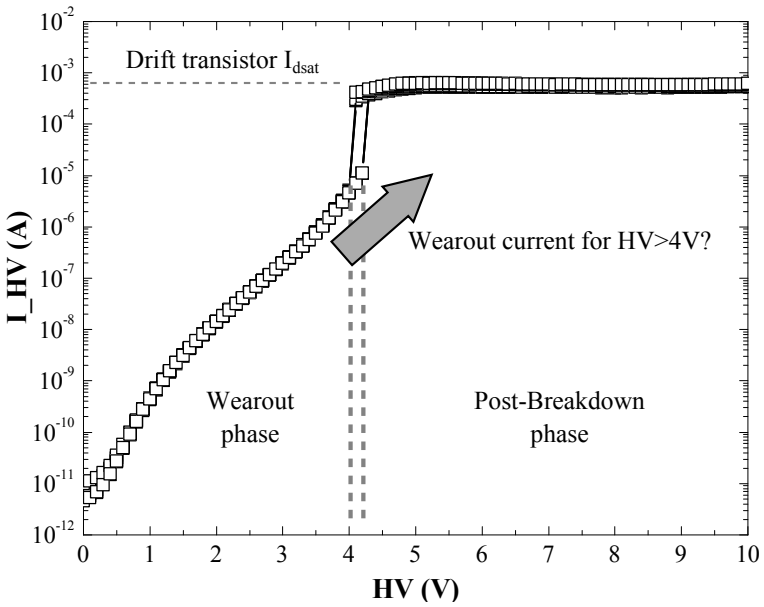


Figure 3.27: Programming current measurement under a DC voltage ramp performed on 6 antifuse bitcells.

The measurement principle is simple. The access transistor is turned on while a DC voltage ramp from 0 to 10V is applied to the HV node of the antifuse bitcell. Thus, the current from HV is measured for each voltage value.

Two voltage domains can be noticed. During the wearout phase from 0 to 4V, the programming current is, in fact, the leakage current from the antifuse capacitor. The gate-oxide breakdown occurs at roughly 4V. Finally the current is steady at the saturation current of the drift transistor in the post-breakdown phase after a steep increase.

DC measurements are particularly useful to evaluate the breakdown voltage of a capacitor. In the case presented in figure 3.27, the breakdown voltage is between 4 and 4.1V. It can be noticed that the behavior of the 6 bitcells is the same except the slight dispersion of the breakdown voltage.

There are several limitations in DC characterization. First, it is obviously not appropriate for TDDB experiments. Second, the wearout phase is limited by the breakdown voltage. In other words it is not possible to measure the wearout current, i.e. the leakage current of the antifuse capacitor, for a voltage higher than the breakdown voltage.

Consequently, additional experimental methodologies are necessary to perform TDDB characterizations and to extend the voltage range of wearout current characterizations.

3.4.2.2 Successive high voltage pulses

Considering a conventional parameter analyzer such as AGILENT 4156c, the current can be measured in a minimum sampling interval of 60μs using a Source Measurement Unit (SMU). The equipment features also Pulse Generator Units (PGU) which offer a more appropriate bandwidth for the characterization of anitufse bitcells. However, the current cannot be measured during a voltage pulse. A solution is therefore needed to perform TDDB characterization by combining a short voltage pulse and accurate current measurements.
Fast gate-oxide degradation characterizations can be performed using successive voltage pulses. A short high voltage pulse is applied to the device. Then, the leakage current is measured under a nominal voltage amplitude. These operations are repeated until hard failure of the capacitor. An algorithm of the experimental method is depicted in figure 3.28.

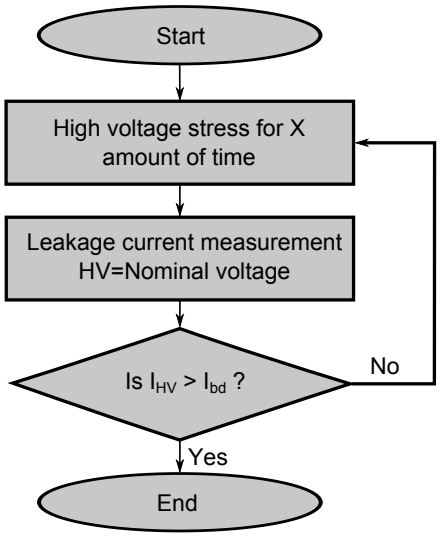
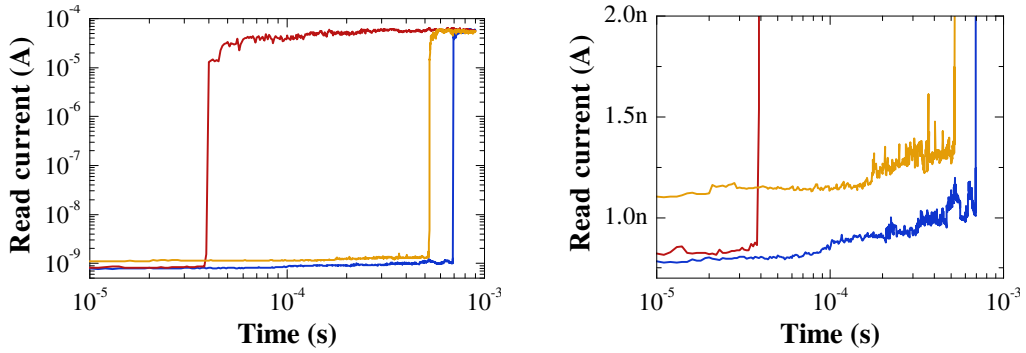


Figure 3.28: Testing flow chart of time dependent dielectric wearout (TDDW) [99].

The cumulative nature of stress-induced damage was reported in literature [99–101]. Thus, voltage pulses can be used to sample the gate-oxide wearout. Nevertheless, the current measurement is performed in a DC mode. Transient aspects are discarded. Measurements were performed on 3 drift antifuse bitcells fabricated in a logic CMOS 45-nm process ($t_{ox} = 17$Å). The access transistor was turned on while high voltage pulses of 1μs duration were applied to the HV node. Thus, the gate current was measured under a nominal voltage condition between two high voltage pulses.



(a) (b)

Figure 3.29: Gate-oxide degradation using successive voltage pulses. HV=4.5V, $T_{pulse} = 1\mu s$, $V_{read} = 1.1V$.

The breakdown event arises after a wearout phase. An increase in noise is noticed before the hard breakdown. By zooming in the wearout phase as in figure 3.29(b), an increase in noise is observed accompanied by a continuous current variation. This type of evolution is defined as progressive in literature (see in section 3.4.1). After the hard breakdown, the maximum current level is reached.

As expected, the breakdown event occurred after a certain time which seems dispersed even on three dice. However, this statement is only qualitative due to the lack of statistical data.

The successive current measurements shown in figure 3.29 can be found with similarities in literature where progressive breakdown is claimed. An increase in noise is clearly noticed with a low amplitude. However, this observation must be counterbalanced with the low voltage amplitude compared to the high programming voltage.

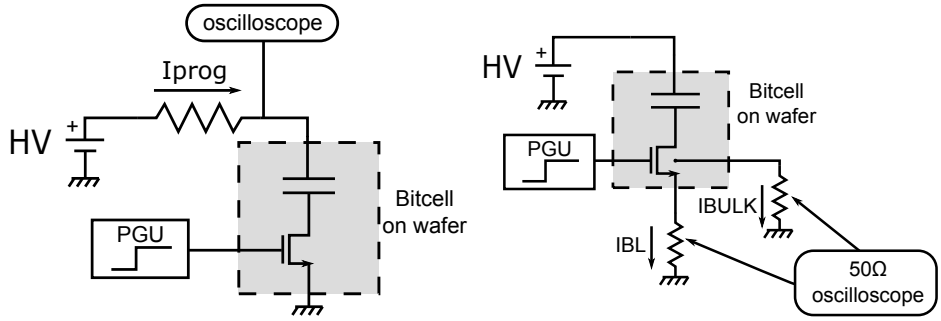
The steep current growth arising at the breakdown event shows that there is still a time window out of the measurement range. In fact, a sampling time of $1\mu s$ is too long. In order to investigate this event deeper, shorter programming voltage pulses are needed. However a pulse width of $1\mu s$ is a limit of the test structure since antifuse bitcells are connected to DC pads.

To conclude, successive voltage pulse gate-oxide degradation experiments brought interesting qualitative results. However, the experimental setup is not satisfying in terms of bandwidth.

3.4.2.3 Current measurements using a series resistor

The limitation of a parameter analyzer for transient measurements is the low bandwidth of SMUs. A pulse generator is more appropriate for this purpose. However,

another equipement is needed to perform current measurements.
A possible setup suitable for transient current measurements is depicted in figure 3.30(a). A capacitor is stressed by a high voltage using a pulse generator while the gate current is measured using a series resistor and an oscilloscope [102, 103].



(a) Discrete series resistor (b) oscilloscope 50-Ω input

Figure 3.30: Experimental setups for transient programming current measurements using series resistors.

Authors reported T_{BD} down to 1μs [102]. The main limitation of this experimental setup is the transient response. The series resistor coupled with parasitic capacitors impact the rise time of the pulse applied on the DUT as well as the voltage measured across the series resistor. However, the measurement accuracy depends on the resistor value: the higher the resistor, the higher the voltage drop. Finally, there is a trade-off accuracy/bandwidth.
The use of a discrete device leads to a significant global cable length and therefore parasitics. This is the reason why this setup was upgraded in the course of the project as depicted in figure 3.30(b). 50-Ω inputs of an oscilloscope are used to measure the bitline and the bulk current during a programming operation. However, the current must be sufficiently high to create a voltage drop in the measurement range of the oscilloscope. Assuming a minimum voltage of 4mV, the minimum current is 40μA. The maximum input voltage is 5V. Therefore, the maximum current is 100mA.
Three main characteristics can be measured using the experimental setups depicted in figure 3.30.

- **Time-to-breakdown:** T_{BD} is reflected by the evolution of the programming current measured across the resistor.

- **Wearout current:** this current flows through the bitcell before the breakdown event.

- **Post-breakdown current:** measurement of transient events and current amplitude.

The three above items can be measured using the setup depicted in figure 3.30(a). However, the parasitics may lead to overlooked transient phenomena.
With the upgraded experimental setup shown in figure 3.30(b), only currents from nodes connected to ground can be measured. Hence, only the bitline and bulk currents are probed. Considering a negligible access transitor gate current, the HV current is: $I_{HV} = I_{BL} + I_{bulk}$. The discrete resistor and the connectors are discarded in figure 3.30(b), thereby reducing parasitics significantly.
An example of programming current measurements is performed using the setup depicted in figure 3.30(b) as shown in figure 3.31. Characterizations were performed on drift antifuse bitcells fabricated in a logic 40-nm CMOS process ($t_{ox} = 17Å$).

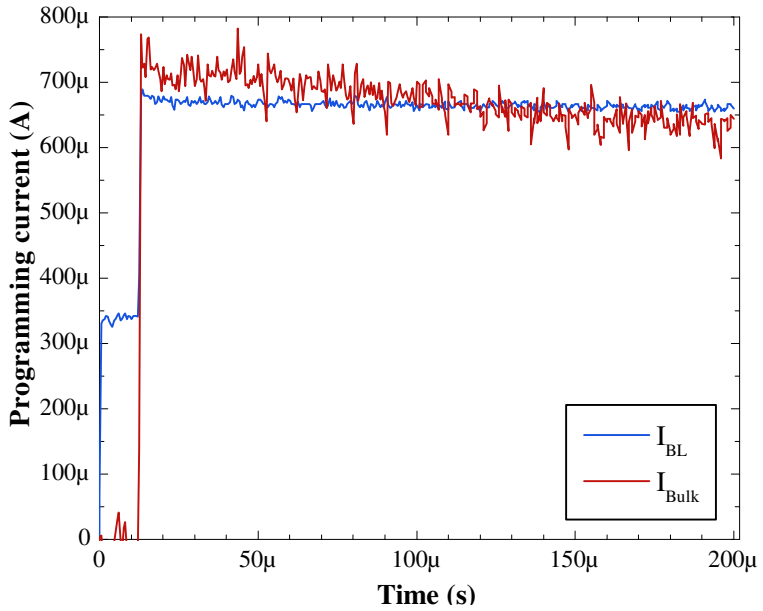


Figure 3.31: Transient programming current measurements performed using series 50-Ω of an oscilloscope. The bitline and the bulk current are measured during a programming operation under HV=5.5V.

A current of about 350μA flows through the bitcell during the wearout phase. This current is measured on the bitline. The bulk current is negligible.
The breakdown is very hard and is shown on the bitline and the bulk current.
In the post-breakdown phase, the bitline current is steady at the saturation of the access transistor (608μA) while the bulk current decays slowly. The particular evolution of the bulk current is discussed in detail in chapter 5.
The minimum measurable T_{BD} range is 1μs using this setup. Shorter transients cannot be measured accurately due to the cables and the DC pads of the test

structure. Nevertheless, the experimental methodology is appropriate to perform characterizations of antifuse bitcells under a suitable programming voltage range although further improvements are needed to reach sub-μs T_{BD}.

3.4.2.4 Transmission Line Pulse

Transmission Line Pulse (TLP) systems are widely used to perform Electrostatic Discharge (ESD) transient characterizations in the Charge Device Model (CDM) time scale. The purpose of this methodology is for example to evaluate the degradation of a dielectric under very short high voltage pulses, e.g. 1ns. A schematic of a standard TLP setup is depicted in figure 3.32.

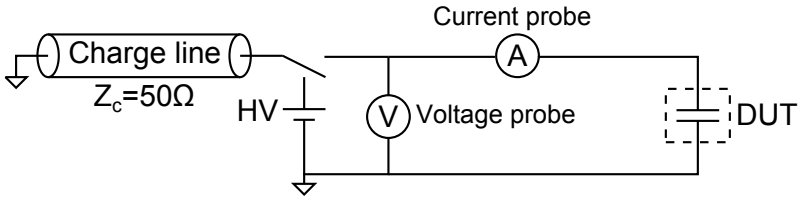
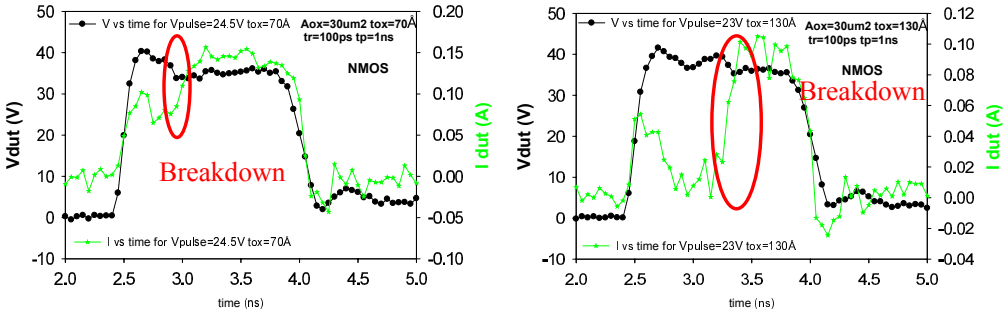


Figure 3.32: Schematic of a Transmission Line Pulse experimental setup [104, 105].

A 50-Ω charge line is charged at HV. Then, a switch is toggled and connects the line to the DUT. The transient current and voltage are measured using an oscilloscope. Examples of breakdown measurements are shown in figure 3.33.



(a) (b)

Figure 3.33: Example of measurements using TLP [104]

In these two examples, the breakdown occurs 1ns after the rising edge of the pulse. Consequently the setup is optimized to generate high voltage pulse for a very short time, e.g. 1ns. Since a particular equipment and test structures are required to perform TLP measurement, none were performed on antifuse bitcells. Besides, the experimental conditions does not seem suitable in terms of voltage and pulse width range.

3.4.2.5 Fast current measurements using a RF bias-Tee

This experimental setup is used to measure current transients occurring during short voltage pulses. In fact this methodology was focused on the characterization of fast charge trapping occuring in high-K dielectrics [106]. The measurement setup is adapted here to an antifuse bitcell as depicted in figure 3.34.

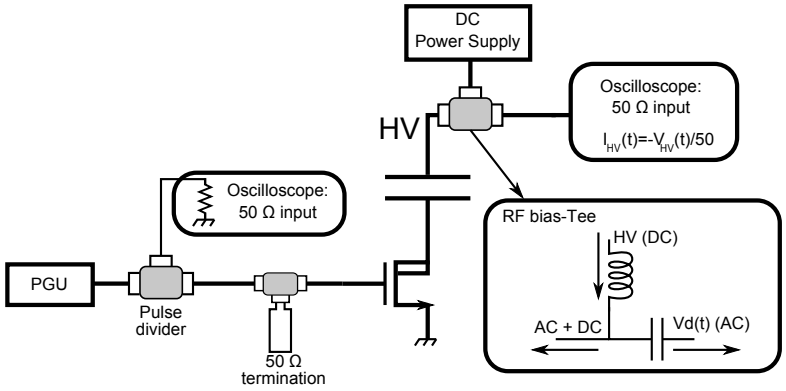


Figure 3.34: RF experimental setup optimized for T_{BD} measurements on antifuse bitcells.

The gate of the capacitor is connected to a RF bias-tee. As shown in inset, a DC voltage is applied to the bitcell using a convential DC source while the AC current is measured using an oscilloscope. In fact the DC and AC parts of the voltage are filtered by the inductor and the capacitor respectively.
The measurement sequence is started by applying a voltage step to the gate of the drift transistor. The scope is trigerred by this driving signal while the evolution of the programming current is measured using the oscilloscope connected to the bias-Tee. Therefore, fast transient current measurements can be performed under a constant voltage stress. This setup is optimized by RF coaxial cables and 50-Ω adapted termination.
Furthermore, the antifuse bitcell is connected to RF pads. A schematic of the test structure is depicted in figure 3.35.

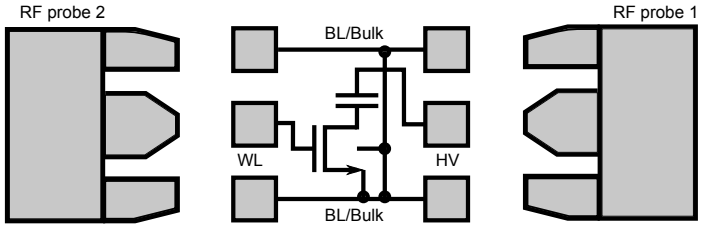


Figure 3.35: Antifuse bitcell connected to a RF PAD.

A RF pad comprises three single pads, one signal pad between two grounded pads. Usually, two structures are used (one in front of another) in order to connect two

different signals. Note that the source and the bulk are connected together to the ground. Consequently, it is not possible to separate the bitline and bulk currents in contrary to the setup presented previously in figure 3.30(b). Only the HV current is measured as the sum of the bitline and bulk current.

An example of TDDB characterization performed on an antifuse bitcell fabricated in the course of the work in a logic 40-nm CMOS process ($t_{ox} = 17$Å) is depicted in figure 3.36.

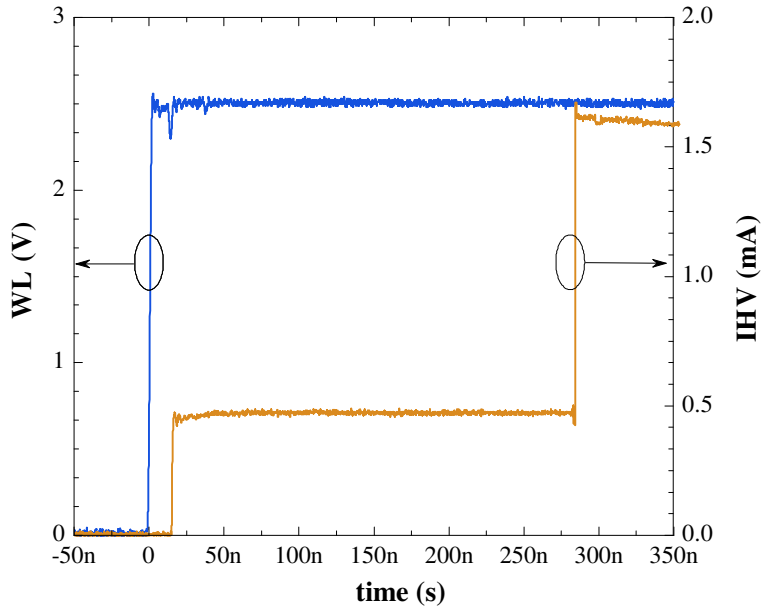


Figure 3.36: Typical transient current evolution during a time-to-breakdown measurement. HV=7V.

The oscilloscope was triggered on WL. Prior to the measurement, a steady DC high voltage of 7V is applied to HV. The gate-oxide, however, is not stressed since the drift transistor is turned off. As seen in previous measurements, a wearout current flows through the bitcell prior to the breakdown event.

After 280ns, a hard breakdown occurs. The HV current reaches 1.6mA because it is the sum of the bitline and the bulk current. Even with the high bandwidth of the RF experimental setup, no noisy phase is observed during the wearout phase. Under such high voltage stress, soft or progressive breakdown does not appear as an issue.

The RF setup features a very high bandwidth and allows sub-μs T_{BD} characterizations.

3.4.3 Conclusion and perspectives

The TDDB characterization of the gate-oxide breakdown occurring during the programming operation of an antifuse bitcell necessitates particular experimental methodologies. There are three main parameters to measure:

- **Time-to-Breakdown:** the bandwidth required to perform TDDB characterization of antifuse bitcells under high voltage can be assumed between 10ns and 100ms. The combination of the RF setup and the series resistor is appropriate to cover this time range.
- **Wearout current:** a conventional DC ramp provides measurement data from 0V to roughly 4V. Measurements using a series resistor is needed to extend the voltage range. Since the bulk current is negligible during this phase, the RF setup is also useful for this purpose, especially for very high voltage. In addition, the wide bandwidth of the RF setup allows the characterization of the breakdown mode, e.g. progressive or hard.
- **Bulk current:** it was shown that the gate-oxide breakdown is also reflected on the bulk current. The separated measurement of the bitline and the bulk current is important to investigate the cause of this current (see in chapter 5).

The different experimental setups presented in this section have been tested, including specific primary test chips, in order to assess the application to the conditions of antifuse bitcell programming. The successive programming voltage pulse methodology introduced in 3.4.2.2 is discarded because the series resistor and the RF setups exhibit better performance and are easier to implement. The transmission line pulse in 3.4.2.4 is also discarded due to a lack of available resources. Indeed, the RF setup allows a sufficient measurement bandwidth.

3.5 Conclusion

The underlying programming physical phenomenon in antifuse bitcells is the breakdown of ultrathin gate-oxide. This topic is discussed in literature mostly regarding reliability issues which corresponds to a slow wearout of the dielectric material stressed under a low voltage. Understanding and modeling the gate-oxide breakdown physics is a major challenge. A strong groundwork for reliability is the statistical study of this failure as the percolation model. It is stated that the degradation of the dielectric is a stochastic process. In other words, defects are randomly generated in the gate-oxide and across the area of a capacitor. Thus, the accumulation

of defects leads to the creation of a percolation path connecting the gate to the substrate.

The gate-oxide breakdown is strongly accelerated by the voltage stress as the higher the stress voltage, the shorter the time-to-breakdown. This property is useful because the degradation of the dielectric is accelerated by stressing capacitors under a voltage higher than nominal conditions. Then, T_{BD} values are extrapolated using a model in order to evaluate the reliability of the gate-oxide under nominal conditions. This topic is actively discussed in literature. Different models based on different physical mechanisms are reported. Even though a formal consensus is lacking, a trend in a power-law voltage dependence of T_{BD} for ultrathin gate-oxide seems the most relevant model so far.

In an antifuse memory, the programming voltage is much higher than the range studied in reliability. Moreover, there are few references about this topic. As mentioned previously, the characteristic T_{BD} value for an antifuse bitcell is 1μs. Therefore a first challenge is to design experimental setups featuring an appropriate bandwidth. Different solutions were found in literature but new developments were needed at die and measurement system level. As a result, a minimum T_{BD} about 10ns can be measured using a RF experimental setup on standard drift antifuse bitcell fabricated in a logic 45-nm CMOS process. Furthermore, the breakdown mode as progressive or hard can be studied. The first measurements performed on antifuse bitcells indicate that a hard failure occurs.

The TDDB modeling of antifuse bitcells programmed under a high voltage is tackled in chapter 4. Characterizations are performed using the experimental setups presented in this chapter. Furthermore, results on high-K/metal gate antifuse bitcells will be presented.

A particular bulk current was detected during antifuse bitcell programming that needs further explanation as detailed in chapter 5.

CHAPTER 4

TDDB modeling for antifuse bitcell design

Understanding the gate-oxide breakdown physics is necessary to optimize the programming operation of antifuse bitcells. There are objectives such as a short programming time, a low programming voltage and a resulting low-ohmic conduction between the electrodes of a broken capacitor. As introduced in chapter 3, the stress voltage has a major impact on the degradation rate of the gate-oxide. The time-dependent dielectric breakdown (TDDB) characterizations and modeling aims at studying the dependence of the time-to-breakdown (T_{BD}) on the stress voltage. Since T_{BD} is one of the parameters which defines the programming time, it makes sense to model the influence of the programming voltage amplitude on the wearout duration towards the best trade-off "short T_{BD}/low HV". The latter trade-off insures a minimal programming energy for a given breakdown effectiveness. The lower the programming power, the easier the effective integration of the charge-pump circuit involved in the programming voltage generation (minimal silicon area).

The knowledge gathered for years regarding the reliability of MOS devices and presented in chapter 3 are proven valid for ultrathin oxide. However, the relevancy of the characterization and modeling methodologies for a high voltage range and short T_{BD} is not studied. Besides, the capacitor area in an antifuse bitcell is much smaller than the typical test structures used to perform reliability assessment of the gate-oxide. A method of TDDB characterization and modeling dedicated to antifuse bitcells is proposed in this chapter.

The methodology is presented in section 4.1. Section 4.2 focuses on the modeling of the wearout current. The modeling of the time-to-breakdown is detailed in section 4.3. An optimization methodology of the bitcell design using a dedicated TDDB

model is proposed in section 4.4. Results from characterizations performed on cascode antifuse bitcell designed and fabricated in a logic 28-nm high-K/metal-gate CMOS process are detailed in section 4.5. Conclusions are drawn in section 4.6.

4.1 Methodology

The drift antifuse bitcell is used in the products of STMicroelectronics for years. This bitcell architecture was presented in section 2.3.2.1 and is characterized in this study. In a logic 45-nm or 40-nm CMOS process, it comprises a thin-oxide $T_{ox} = 17\,\mathring{A}$ capacitor connected in series with a thick-oxide $T_{ox} = 31\,\mathring{A}$ or $50\,\mathring{A}$ drift transistor. Schematic and cross-sectional views are depicted in figure 4.1.

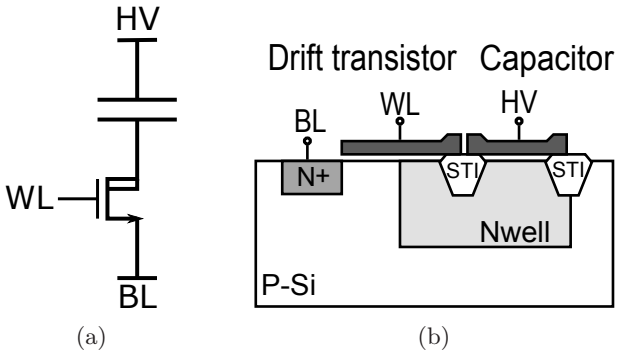


Figure 4.1: Schematic and cross section of the drift antifuse bitcell.

The measurement of T_{BD} is performed by means of a high voltage applied to the HV node and by turning on the drift transistor.

4.1.1 Typical Time-to-Breakdown measurements

The programming current can be recorded using either series resistors as in the experimental setup detailed in 3.4.2.3 and depicted in figure 3.30 or an RF bias-Tee for better performance as presented in 3.4.2.5.
Examples of T_{BD} measurements are shown in figure 4.2. Only the bitline current is displayed. The issue regarding the bulk current overshoot briefly presented in chapter 3 is discussed in details in chapter 5.

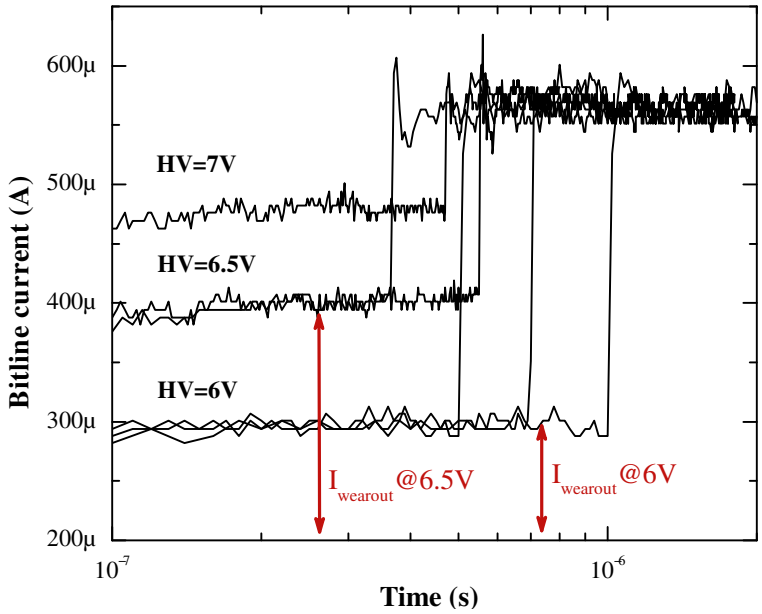


Figure 4.2: Bitline current waveform for a drift antifuse bitcell designed and fabricated in a logic CMOS 40-nm process. (BL=0V, WL=2.5V for $T_{ox} = 50$ Å)

As presented in chapter 3, the time-to-breakdown of antifuse bitcells obtained for a programming voltage range from 6 to 7V is shorter than 1μs. A significant wearout current can be noticed on the measurements plotted in figure 4.2. Since the saturation current of the drift transistor is measured at 580μA, it comes that the transistor operates in the ohmic regime during the wearout phase. As a consequence, the drain-to-source voltage must be evaluated and taken into account in order to determine accurately the voltage across the capacitor.

4.1.2 Wearout current and voltage operating point

The gate-oxide breakdown is a phenomenon strongly accelerated by the stress voltage HV. Consequently, the voltage across the antifuse capacitor must be maximized in order to minimize T_{BD}. The accurate measurement and modeling of the wearout current appears therefore essential in order to evaluate the voltage operating point of an antifuse bitcell during the wearout phase.
The leakage current from an antifuse capacitor can be measured by means of applying a voltage ramp on the HV electrodes. An example of characterization was shown in chapter 3, the measurements are plotted again in figure 4.3 for convenience.

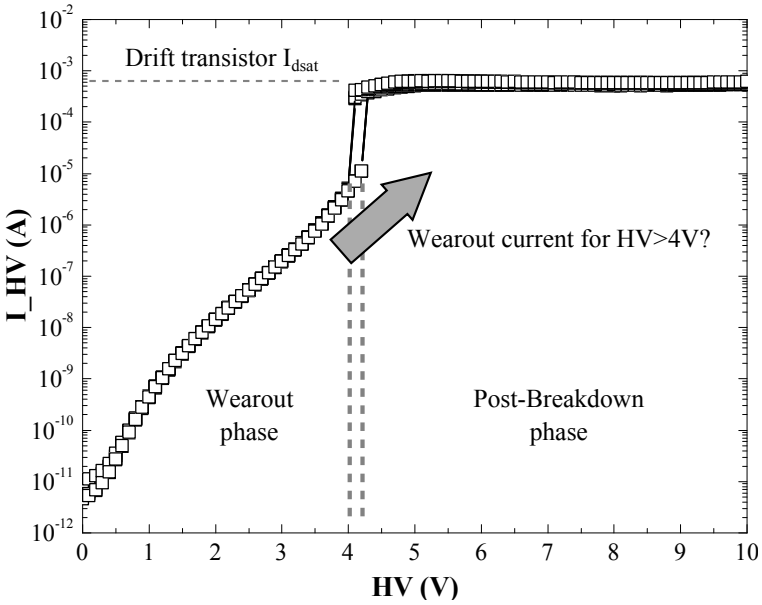


Figure 4.3: Programming current measurement under a DC voltage ramp performed on 6 antifuse bitcells.

The wearout current is measured accurately as a function of the HV amplitude. However, the results obtained for a voltage higher than the breakdown voltage do not characterize a wearout current due to the equipment integration time (few ms). Wearout current measurements are needed for a voltage range up to 7V (programming voltage of 40-nm CMOS antifuse memory products). The DC measurements are therefore not sufficient. This limitation is overcome using a transient experimental setup (see in chapter 3, section 3.4.2). Since the current is constant during the wearout phase (see in figure 4.2), an average value can be extracted from transient measurements before the breakdown event. Thus, the DC characteristic of the antifuse capacitor leakage current is completed by transient measurements as plotted in figure 4.4.

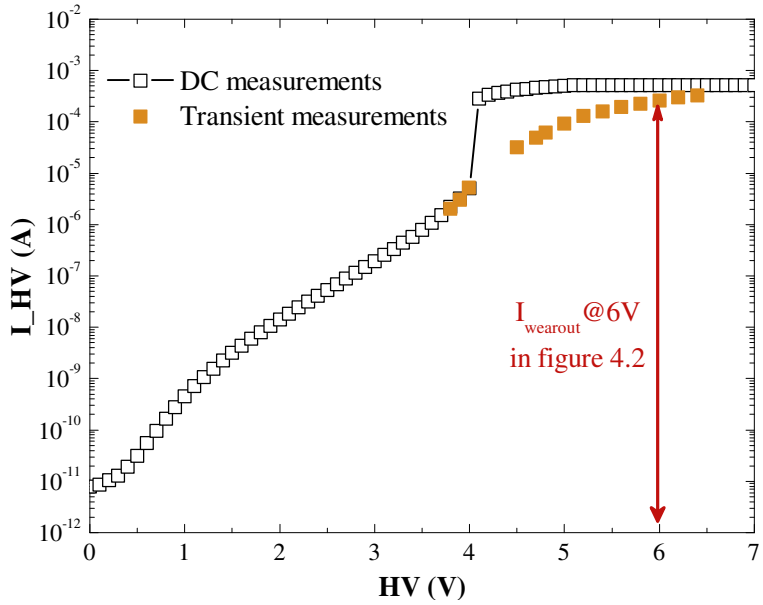


Figure 4.4: DC and transient wearout current measurements performed on antifuse bitcells designed and fabricated in a logic 40-nm CMOS process.

The transient measurements are successfully appended to DC measurements. The relevancy of the experimental methodology (see in chapter 3, section 3.4.2) is therefore demonstrated. It was experimentally shown that the wearout current is similar from one bitcell to another. The DC characteristics are perfectly superimposed and the successive transient measurement data appear continuous. The reason of this reproducibility is that the thin gate-oxide is accurately controlled during the fabrication process. Due to the stringent requirements in term of defect density, the leakage current is well controlled and is a sensitive parameter regarding the performance of complex SoCs.

The current transport process in the ultrathin gate-oxide of the antifuse capacitor is investigated in the next section. The mechanisms presented in chapter 3 are driven by the electric field applied across the dielectric. The accurate determination of the voltage across the capacitor is therefore important to achieve a proper modeling of the wearout current.

The voltage drop across the drift transistor can be determined by plotting the wearout current characteristic of an antifuse capacitor and the $I_{Drain} - V_{Drain}$ characteristic of the drift transistor such that the voltage across both devices is determined at the intersection of the two curves. An example of operating point calculations for a bitcell designed and fabricated in a logic 45-nm CMOS process is shown in figure 4.5 for HV=5 and 6V respectively.

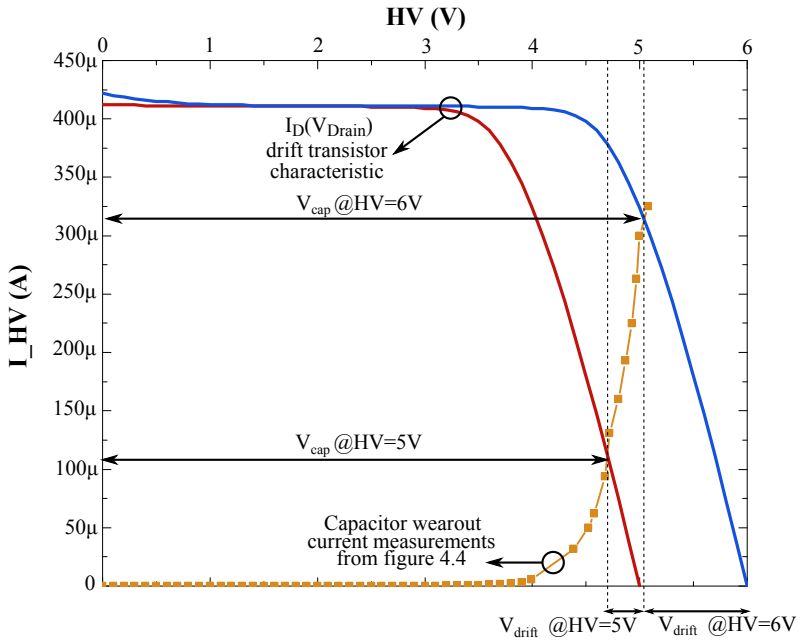


Figure 4.5: Calculation of voltage operating points for HV=5 and 6V in a drift antifuse bitcell fabricated in a logic 45-nm CMOS process.

The wearout current amplitudes for HV=5 and 6V are respectively 120μA and 320μA. In this example, the on-state resistance of the drift transistor is nearly 2800Ω. Hence, the voltage drop across the drift transistor is 336mV for HV=5V and 896mV for HV=6V respectively. Even though HV is increased by 1V, the capacitor voltage is only increased by 440mV.

The impact of the wearout current amplitude and the on-state resistance of the drift transistor is clearly emphasized in this practical example. Since the programming voltage is generated by a charge-pump circuit, increasing its amplitude by 1V leads to a significant cost in terms of circuit area and power consumption. Furthermore, the effectiveness is limited due to the series device. The modeling of the time-to-breakdown and the wearout current is, indeed, relevant if the effective capacitor voltage is accurately determined. This approach is further illustrated in figure 4.6.

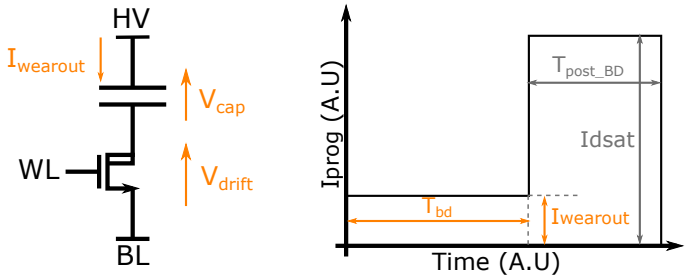


Figure 4.6: Simplified programming operation of an antifuse bitcell. The characterization and modeling methodologies are focused on T_{BD} and $I_{wearout}$.

The contributions of different parameters such as the capacitor area and the dimensions of the access transistor are thoroughly studied in this chapter in order to minimize the time-to-breakdown and the programming voltage amplitude. Furthermore, the reduction of the programming voltage enables the downsizing of the charge-pump circuit. There is also an interest to understand the contribution and the duration of the wearout phase and the post-breakdown phase. The latter phase is detailed in chapter 5.

4.2 Wearout current modeling

The wearout current flows through the gate-oxide during the programming operation and causes damages leading to the breakdown of the dielectric material. Since this current leads to a significant voltage drop across the access transistor, a model is needed in order to accurately determined the effective voltage across the capacitor.

4.2.1 Fowler-Nordheim tunneling

As presented in chapter 3, there are several (possible current) transport processes which model a leakage current through a dielectric. Since the thin gate-oxide in a fully silicon CMOS process such as 45-nm and 40-nm is accurately controlled, a very low defect density can be assumed. Hence, the current transport should be driven by the barrier heights of the electrodes. A Fowler-Nordheim tunneling mechanism can be assumed due to the high programming voltage amplitude.

4.2.1.1 Analytical expression

The equation of the current density in a Fowler-Nordheim conduction mode was given in 3.2.1.1 as:

$$J_{ox} = C \cdot E_{ox}^2 \cdot \exp\left(-\frac{D}{E_{ox}}\right) \tag{4.1}$$

In order to extract the parameters C and D, equation (4.1) is classicaly transformed as:

$$\ln\left(\frac{J_{ox}}{E_{ox}^2}\right) = \ln(C) - \frac{D}{E_{ox}} \tag{4.2}$$

Plotting $\ln\left(\frac{\mathrm{J_{ox}}}{\mathrm{E_{ox}^2}}\right)$ versus $\frac{1}{\mathrm{E_{ox}}}$, the range of $\mathrm{E_{ox}}$ in which the FN model is relevant should be reflected by a straight line allowing the identification of $ln(C)$ and D.

4.2.1.2 Practical example

Two drift antifuse bitcells designed and fabricated in a logic 45-nm CMOS process were used in order to verify the relevancy of a Fowler-Nordheim tunneling as a wearout current model. Dimensions are given in table 4.1.

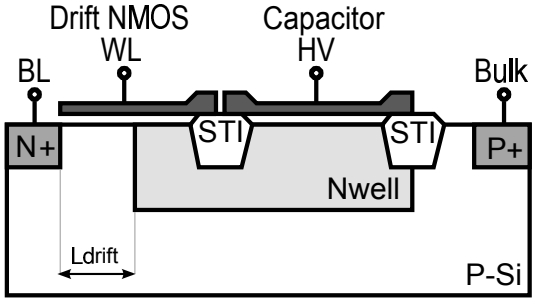


Figure 4.7: Cross section of a drift antifuse bitcell.

CMOS	$\mathbf{A_{bitcell}}$	$\mathbf{A_{cap}}$	$\mathbf{W_{drift}}$	$\mathbf{L_{drift}}$
45nm	$16\mu m^2$	$0.97\mu m^2$	$2.2\mu m$	$0.85\mu m$
45nm	$5\mu m^2$	$0.13\mu m^2$	$0.5\mu m$	$0.25\mu m$

Table 4.1: Dimensions of antifuse bitcells as designed and fabricated 45-nm CMOS process.

The devices exhibit different antifuse capacitor and drift transistor dimensions. They are denoted by their area for convenience as 16-μm^2 and 5-μm^2 bitcells in the following case-study.

To achieve the modeling of the wearout current, DC and transient characterizations are performed for a HV amplitude up to 6.4V [1]. As mentioned previously, the voltage drop across the drift transistor is critical in the high voltage domain resulting to a significant wearout current. Thus, the effective capacitor voltage was accurately determined for each current measurement.

[1] Limitations of the experimental measurement system forbid to achieve a higher voltage.

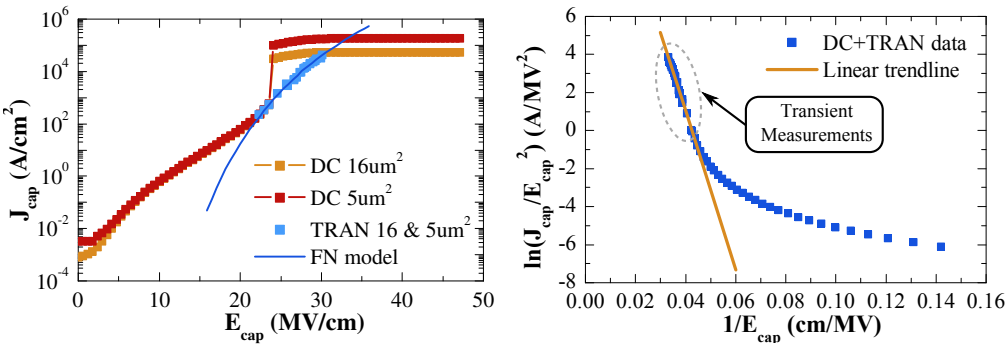
DC and transient measurement data are plotted in figure 4.8(a) as the current density J_{cap} versus the electric field E_{cap}. Both parameters are calculated as:

$$J_{cap} = \frac{I_{cap}}{A_{cap}} \quad \text{and} \quad E_{cap} = \frac{V_{cap}}{t_{ox}} \tag{4.3}$$

The definition of E_{cap} can be questioned. The calculation of the electric field value applied across the gate-oxide is slightly more complex. An expression of E_{ox} is given in the following equation [107]:

$$E_{ox} = \frac{V_{cap} - V_{FB} - R_s \cdot I_{cap}}{t_{ox}} \tag{4.4}$$

where V_{cap} is the voltage across the capacitor, V_{FB} is the flat band voltage, R_s is the series resistance in the structure and I is the tunnel current. Equation (4.4) shows that there are voltage drops to be taken into account in order to evaluate the electric field across the gate-oxide. In the case of the drift antifuse bitcell, the term $R_s \cdot I_{cap}$ is taken into account in the correction of the voltage drop in the drift transistor. The flat-band voltage is neglected ($\approx$ 1V [107]) in a first approach.
As shown in section 4.1, the DC and transient measurements are correctly appended. Those data are then plotted in figure 4.8(b) as defined in equation 4.2.



(a) (b)

Figure 4.8: (a) DC and transient current densities for 16-μm^2 and 5-μm^2 bitcells respectively and plotted versus E_{cap} and fitted with the Fowler-Nordeim model. (b) Identification of the Fowler-Nordheim parameters using a linear trendline.

As expected, the Fowler-Nordheim conduction mode is reflected by a straight line allowing the extraction of the parameters C and D. Finally, the Fowler-Nordheim model in equation (4.1) fits correctly the measurements as shown in figure 4.8(a). The Fowler-Nordheim current transport process is relevant to model a tunnel current under high voltage because the carriers cross a triangle barrier due to the severe band bending. It is commonplace to assume that the boundary between direct and Fowler-Nordheim conduction mode is about 5MV/cm [108]. However, it is noticed

that the correct fitting in figure 4.8(b) appears for an electric field higher than 25MV/cm. The parameters neglected in the calculation of the effective electric field applied across the capacitor may lead to an erroneous range in which the identified is correctly fitted. The flat-band voltage was indeed not taken into account. However, a drop of 1V yields to a drop of 5.9MV/cm according to (4.4). The depletion of the polysilicon gate is also a parameter which lead to a voltage drop and should be accounted in the calculation of the electric field. The identification of a physical parameter is necessary to conclude on the nature of the current transport process involved in the wearout phase. The parameters C and D were defined in chapter 3, equation (3.2). The identification of the injecting electrode barrier height Φ_{b} would be a relevant evidence regarding the validity of a Fowler-Nordheim model.
Even though assumptions must be verified to formally conclude on the pertinence of a Fowler-Nordheim model, the expression defined in equation (4.1) is used to model the wearout current. The latter equation is derived for different capacitor areas and as a function of the capacitor voltage as follows:

$$I_{FN} = A_{cap} \cdot C \cdot \left(\frac{V_{ox}}{t_{ox}}\right)^2 \cdot \exp\left(-D \cdot \frac{t_{ox}}{V_{ox}}\right) \tag{4.5}$$

The values of parameters C and D are the same as calculated using equation (4.2). Fowler-Nordheim-like models were then derived for each bitcell according to the capacitor area given in table 4.1. DC and transient wearout current measurements are plotted in figure 4.9 with the corresponding models.

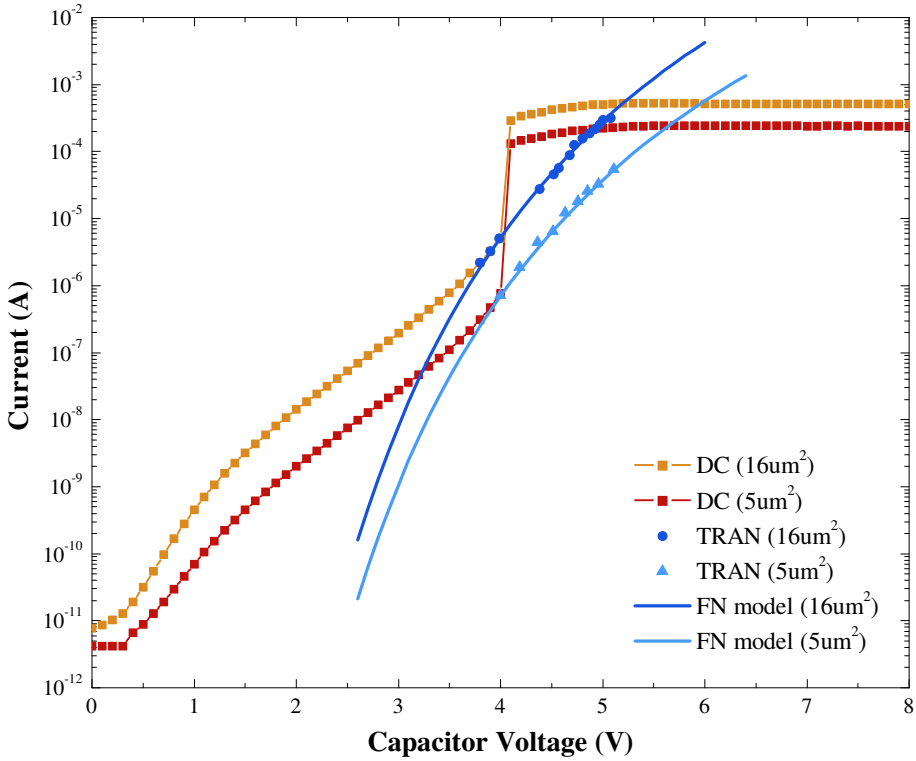


Figure 4.9: Fowler-Nordheim-like wearout current models fitted with DC and transient measurements for 16-μm^2 and 5-μm^2 bitcells.

The influence of the capacitor area is clearly emphasized as the larger the area the higher the wearout current. The models are correctly fitted with the measurement data, thereby demonstrating the pertinence of the methodology. The parameters C and D can be identified experimentally on a reference device. Then, a Fowler-Nordheim-like model can be derived for different capacitor area.

The dependence on temperature of a tunnel current was discussed in chapter 3. Since the Direct and Fowler-Nordheim conduction modes are driven by the voltage bias and the barrier height of the electrodes, the dependence on temperature is negligible whereas trap-assisted mechanisms such as the Frenkel-pool transport are strongly activated by this parameter. This property was emphasized in section 3.2.1.3, equation (3.5). Consequently, the characterization of the wearout current for different temperatures is a solution to verify one property or another. DC measurements were performed on antifuse bitcells designed and fabricated in a logic 40-nm CMOS process under 25°C, 80°C and 125°C. Results are plotted in figure 4.10.

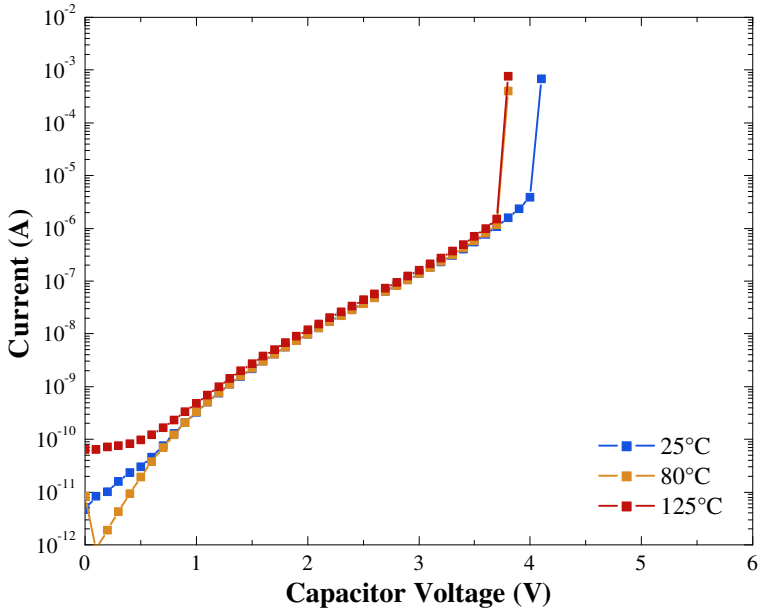


Figure 4.10: DC measurements under 25°C, 80°C and 125°C performed on antifuse bitcells designed and fabricated in a logic 40-nm CMOS process.

The wearout current is similar from 1 to 4V. Since the current transport process is not activated by the temperature, a trap-assisted conduction mode is not relevant or negligible due to the low defect density in ultrathin gate-oxide.

4.2.2 Conclusion

The modeling of the wearout current as a Fowler-Nordheim-like conduction is a first achievement in the characterization of the wearout phase. Indeed, the current can be calculated in a high voltage range and for different capacitor areas. Thus, the impact of the antifuse capacitor dimensions on the operating point can be estimated. Nevertheless, measurements must be performed on a reference device in order to identify the Fowler-Nordheim parameters. The characterization of the on-state resistance of the access device is also essential. Extra work is needed in this field in order to formally identified to the current transport process. In this respect, the identification of the injecting electrode barrier height is a relevant perspective.

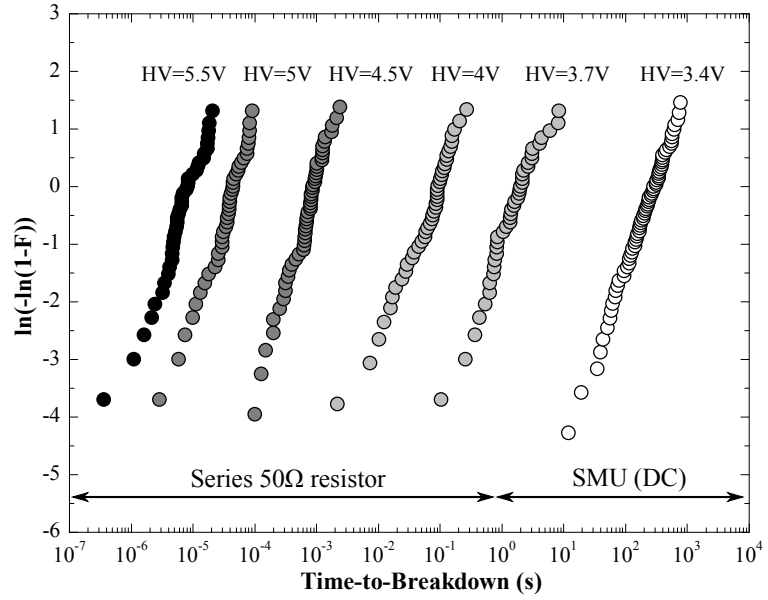
Furthermore, DC characterizations performed under different temperature conditions have shown that the wearout current was not impacted. Trap-assisted transport processes can be therefore discarded for the modeling of the wearout current.

4.3 Time-to-breakdown modeling

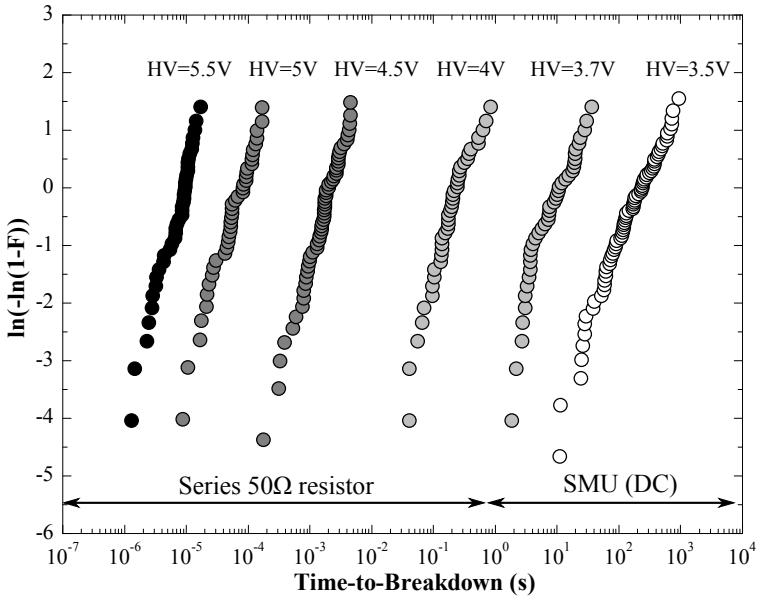
The gate-oxide breakdown is a physical phenomenon mainly accelerated by the stress voltage applied to the device under test. This acceleration is characterized by the time-to-breakdown which describes the duration of the wearout phase. Like for the modeling of the wearout current, the effective voltage across the capacitor must be accurately determined in order to achieve a relevant modeling of the time-to breakdown voltage acceleration. The methodology and the results are presented in this section.

4.3.1 Measurements and distributions

Time-to-breakdown measurements were performed on 16-μm^2 and 5-μm^2 bitcells for a programming voltage range from 3.4 to 5.5V. Most of the measurements were performed using the experimental setup depicted in chapter 3, figure 3.30(b). The bitline current is measured using a 50Ω input of an oscilloscope. However, a conventional Source Measurement Unit (SMU) was used for T_{BD} longer than 1s as experimented in reliability. Distributions of T_{BD} for 16-μm^2 and 5-μm^2 bitcells are plotted in figure 4.11.



(a)



(b)

Figure 4.11: T_{BD} distributions for 16-μm^2 (a) and 5-μm^2 (b) bitcells.

Each distribution is a set of 50 bitcells. A range of 10 orders of magnitude of T_{BD} is covered from 300ns to 1000s. The Weibull slope β is constant and slightly larger than 1 as commonly observed for ultrathin gate-oxides. The experimental setup is therefore relevant for the characterization of short T_{BD} on antifuse bitcells. The identified Weibull slopes (β) for the 16-μm^2 and 5-μm^2 bitcells are listed in table 4.2.

HV (V)	3.4	3.5	3.7	4	4.5	5	5.5
$\beta_{16\mu m^2}$	1.35	-	1.26	1.16	1.67	1.47	1.37
$\beta_{5\mu m^2}$	-	1.18	1.39	1.63	1.70	1.58	1.87

Table 4.2: Weibull slopes of the T_{BD} distributions for 16-μm^2 and 5-μm^2 bitcells respectively (45-nm CMOS).

4.3.2 Identification of the voltage-acceleration law

Different voltage-acceleration law of T_{BD} were presented in chapter 3, section 3.2.3. It was concluded that the power law is a relevant model for ultrathin gate-oxide breakdown as defined in the following equation:

$$T_{BD} \propto \frac{1}{A_{cap}} \cdot V_{cap}^{n} \tag{4.6}$$

This equation is, in fact, the combination of the area-scaling property of the percolation model and the conventional power law. Equation (4.6) emphasizes that the larger the capacitor area and the higher the capacitor voltage, the shorter the T_{BD} value. The pertinence of this equation in a high voltage range and thereby short T_{BD} was not proven beforehand. Thus, 50% T_{BD} values from the distributions plotted in figure 4.11 are plotted versus the capacitor voltage in figure 4.12, which is determined using prior wearout current measurements. Then, a power law is identified according to measurement data for the 16-μm^2 and 5-μm^2 bitcells respectively.

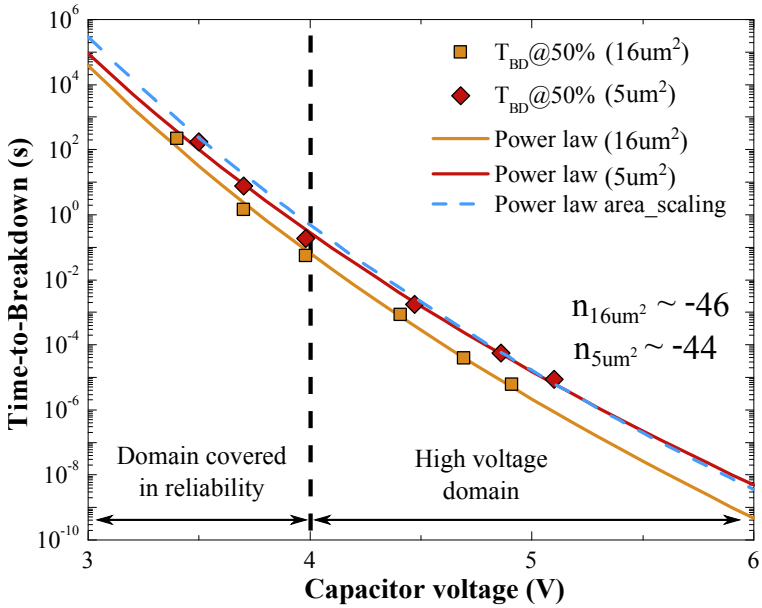


Figure 4.12: T_{BD} plotted versus V_{cap} and fitted against measurements.

Power-law models are correctly fitted against measurements for both bitcells. Since the validity was proven up to 4V, the domain is now extended up to 5V with a corresponding T_{BD} of 1μs. The influence of the capacitor area is also clearly seen. The acceleration is somewhat similar for both bitcell while the 16-μm^2 bitcell exhibit shorter T_{BD} values due to a larger capacitor area than the 5-μm^2 bitcell.
The area-scaling property is verified by projecting the power-law model of the 16-μm^2 bitcell to the capacitor area of the 5-μm^2 bitcell as defined in equation 4.7.

$$T_{BD\_scaled} = \alpha_{16\mu m^2} \cdot \frac{A_{cap\_16\mu m^2}}{A_{cap\_5\mu m^2}} \cdot V_{cap}^{n_{16\mu m^2}} \tag{4.7}$$

$\alpha_{16\mu m^2}$ is the proportionality factor and $n_{16\mu m^2}$ is the acceleration factor identified from the T_{BD} power-law model of the 16-μm^2 bitcell. The projection is achieved by adding a factor as the ratio of the capacitor areas of the reference 16-μm^2 bitcell and the 5-μm^2 bitcell. As a result, the projected model is plotted in figure 4.12.
The area-scaling property appears relevant. The discrepancy with the measurements comes from the difference between the two acceleration factors identified from the power-law models. Although the error on the projections from one device area to another or in a high voltage range may be significant, the accuracy is sufficient to identify major trends.

4.3.3 RF measurements

The modeling of the time-to-breakdown using a power law appeared consistent in a programming voltage range up to 5.5V. However, measurements under higher voltage amplitudes and shorter time-to-breakdown are needed to cover the practical range up to 7V as used in industrial products. This limitation comes from the bandwidth of the experimental setup. Indeed the measurement of the bitline current is not relevant for a mean T_{BD} value shorter than 1μs. Very short T_{BD} values about hundreds of nanosecond can be tolerated as plotted in figure 4.11(a).
To overcome this limitation in bandwidth, TDDB characterizations were performed using a RF experimental setup. Thus, measurements are performed in a programming voltage up to 7V. As presented in chapter 3, section 3.4.2.5, a dedicated test vehicle embedding the bitcell and RF pads is required. Such test structures were not available in a logic 45-nm CMOS process. Therefore, the following reported measurements were performed on antifuse bitcell designed and fabricated in a logic 40-nm CMOS process. Table 4.3 gives the dimensions of the tested device.

CMOS	$A_{bitcell}$	A_{cap}	W_{drift}	L_{drift}
40nm	$10\mu m^2$	$0.58\mu m^2$	$2\mu m$	$0.3\mu m$

Table 4.3: Dimensions of an antifuse bitcell as designed and fabricated in a logic 40-nm CMOS process.

TDDB characterizations were performed in a programming voltage range from 4 to 7V. The RF setup was used to perform T_{BD} measurements for HV=6 and 7V while characterizations in a lower range were achieved using a series 50Ω.
T_{BD} distributions are plotted in figure 4.13

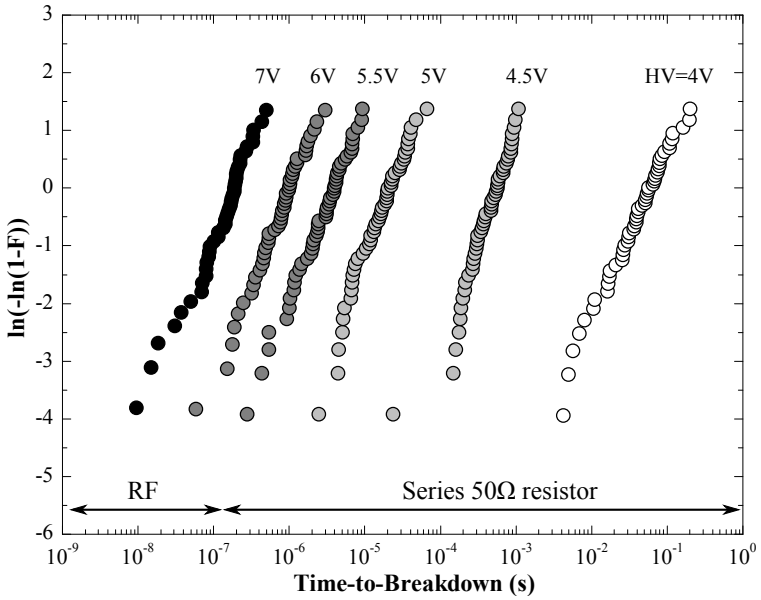


Figure 4.13: T_{BD} distributions for 10-μm^2 bitcells designed and fabricated in a logic 40-nm CMOS process.

The use of the RF experimental setup is successful to perform T_{BD} measurements shorter than 10ns. Thus, distributions for high programming voltage used in antifuse memory products are plotted. The Weibull slope is not changed even at HV=7V. The RF experimental setup is therefore relevant to perform TDDB characterizations on antifuse bitcells. The slopes β are given in table 4.4. The values are similar with the ones identified on 45-nm antifuse bitcells in table 4.2.

HV (V)	4	4.5	5	5.5	6	7
$\beta_{10\mu m^2}$	1.26	1.66	1.56	1.39	1.38	1.35

Table 4.4: Weibull slopes of the T_{BD} distributions for 10-μm^2 bitcells (40-nm CMOS).

Voltage-acceleration law The mean value identified from the Weibull distributions are plotted versus the capacitor voltage in figure 4.14. The voltage across the different series elements are corrected.

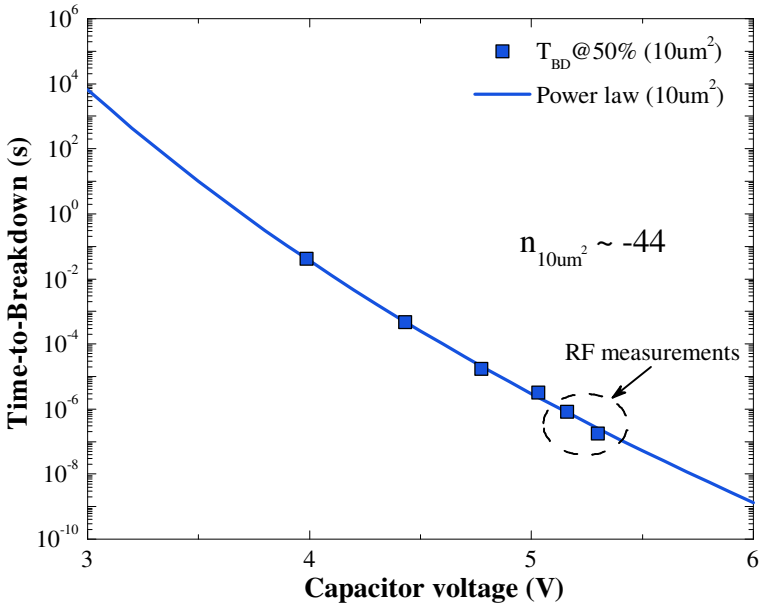


Figure 4.14: T_{BD} plotted versus V_{cap} and fitted against measurements for 10-μm^2 bitcells designed and fabricated in a logic 40-nm CMOS process.

A power-law model was identified from the T_{BD} measurements. The acceleration factor is in the range reported in literature (-43 ± 3). The measurements performed using the RF experimental appears therefore relevant to identify a power-law model in a high voltage range and consolidate the present approach.

4.3.4 Conclusion

The time-to-breakdown measurements performed either using a 50Ω series resistor or a RF bias-Tee appeared relevant. T_{BD} distributions for a corresponding stress voltage up to 7V were successfully plotted on a Weibull scale and exhibited the same property as distributions of measurements performed in a low voltage range. Indeed, the slope β was not impacted by the antifuse bitcell architecture or the experimental setup.

The voltage-acceleration of the time-to-breakdown was correctly modeled by a power law. However, this approach is relevant if the capacitor voltage is accurately determined. In other words, the voltage drop across the different series elements such as the drift transistor and the measurement resistor are taken into account. This is the reason why the wearout current amplitude is a key parameter in accurate model-

ing of T_{BD}. Even for a programming voltage up to 7V, T_{BD} (V_{cap}) was successfully modeled by a power law and correctly fitted with measurements.
The identification of a physical mechanism leading to a power-law dependence of T_{BD} is valuable in order to build up an accurate model and can be considered as a relevant future work. For now, this study is focused on the identification of the main trends on the dependence of T_{BD} on the programming voltage amplitude and the dimensions of the antifuse bitcell.

4.4 Optimization of the antifuse bitcell design

The study of the modeling of the wearout current and the time-to-breakdown described in sections 4.2 and 4.3 emphasized the contributions of the capacitor area and the drift transistor during the wearout phase. A modeling approach and a method of optimization are proposed in this section in order to minimize the programming voltage amplitude, the time-to-breakdown and the dimensions of the antifuse bitcell.

4.4.1 Modeling approach

The definition of the model is described in the following section. An equivalent circuit to an antifuse bitcell during the wearout phase is proposed in order to derive an equation of the capacitor voltage as the function of the programming voltage, the capacitor area and the dimensions of the drift transistor. Then, the impact of those parameters on the time-to-breakdown and the wearout current can be studied.

4.4.1.1 Antifuse bitcell equivalent circuit

During the wearout phase, the high programming voltage applied to the gate of the antifuse capacitor leads to the flow of a Fowler-Nordheim tunnel current through the dielectric and the drift transistor. Since the wearout current is lower than the saturation current of the access device, the drift transistor operates in the linear regime. The drain-to-source voltage can be therefore calculated by the product of the wearout current amplitude and the on-state resistance (R_{on}) of this device. An equivalent circuit of an antifuse bitcell is depicted in figure 4.15 based on the latter assumption.

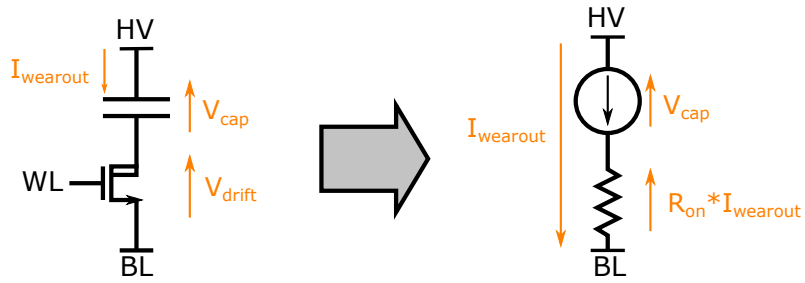


Figure 4.15: Modeling of an antifuse bitcell during the wearout phase using a current source and a series resistor.

The modeling of the Fowler-Nordeim wearout current and the voltage acceleration of the time-to-breakdown is relevant if the effective voltage amplitude across the capacitor is accurately determined. According to the circuit depicted in figure 4.15 a simple equation of the capacitor voltage can be derived as a function of the programming voltage amplitude, the on-state resistance (R_{on}) of the access transistor and the wearout current.

$$V_{cap} = HV - R_{on} \cdot I_{wearout} \tag{4.8}$$

The wearout current is defined using the Fowler-Nordheim model as previously defined in equation (4.5). The expression is recalled for clarity.

$$I_{wearout} = A_{cap} \cdot C \cdot \left(\frac{V_{cap}}{t_{ox}}\right)^2 \cdot \exp\left(-D \cdot \frac{t_{ox}}{V_{cap}}\right) \tag{4.9}$$

The expression of the on-state resistance of a MOS transistor operating in the linear regime is [109]:

$$R_{on} = \frac{L}{\mu_n C_{ox} W \left(V_{gs} - V_{th}\right)} \tag{4.10}$$

4.4.1.2 Expression of V_{cap}

Assuming that the Fowler-Nordheim parameters C and D, the gate-oxide thickness, the capacitor, the dimensions and the process parameters (μ_n, C_{ox} and V_{th}) of the drift transistor are identified, the capacitor voltage V_{cap} can be calculated as a function of the programming voltage amplitude HV using equations (4.8), (4.9) and (4.10).

Then, a solver is used to compute V_{cap} as a function of HV using equation (4.8). However, this modeling approach is relevant if the drift transistor operates in a triode regime and if the wearout current is driven by a Fowler-Nordheim transport process.

4.4.1.3 Model output

Once the capacitor voltage has been computed, the model can yield different results. The power-law was introduced previously as a relevant model of time-to-breakdown for ultrathin oxide capacitor. An equation of T_{BD} was given in (4.6). The proportionality factor and the voltage-acceleration factor n in (4.6) must be identified from measurements. Thus, the time-to-breakdown can be computed and plotted versus V_{cap} and HV.
A similar result can be obtained for the wearout current as $I_{wearout}(V_{cap}, HV)$.

To conclude, the modeling methodology is summarized in figure 4.16.

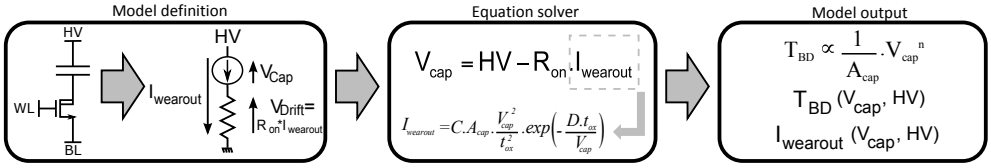


Figure 4.16: Modeling approach of $I_{wearout}$ and T_{BD}. V_{cap} is determined using an equation solver as a function of HV, the on-state resistance of the drift transistor and the area of the antifuse capacitor. As results, $I_{wearout}$ and T_{BD} are plotted versus V_{cap} and HV.

4.4.2 Application and verification

The modeling approach presented previously is verified on a practical case-study in order to verify the pertinence of the proposed methodology.

4.4.2.1 Antifuse bitcell dimensions

Three different drift bitcell architectures were designed and fabricated in a logic 40-nm CMOS process in order to apply the modeling method. Since the study focuses on the voltage operating point, the bitcells feature different capacitor areas whereas the dimensions of the drift transistor are unchanged as given in table 4.5.

Bitcell	$\mathbf{A_{cap}}$	$\mathbf{W_{drift}}$	$\mathbf{L_{drift}}$
Small	$0.15\mu m^2$	$2\mu m$	$0.3\mu m$
Medium	$0.58\mu m^2$	$2\mu m$	$0.3\mu m$
Large	$2.38\mu m^2$	$2\mu m$	$0.3\mu m$

Table 4.5: Dimensions of drift antifuse bitcells as designed and fabricated in a logic 40-nm CMOS process.

4.4.2.2 Identification of parameters

Time-to-breakdown measurements were performed on the three antifuse bitcells for a programming voltage range from 4 to 5.5V (sets of 50 samples). In addition, the medium bitcell was available in a RF test structure. The voltage range was therefore extended up to 7V for this device. Some measurements were presented previously in figures 4.13 and 4.14. T_{BD} distributions of the small, medium and large bitcells are plotted in figure 4.17.

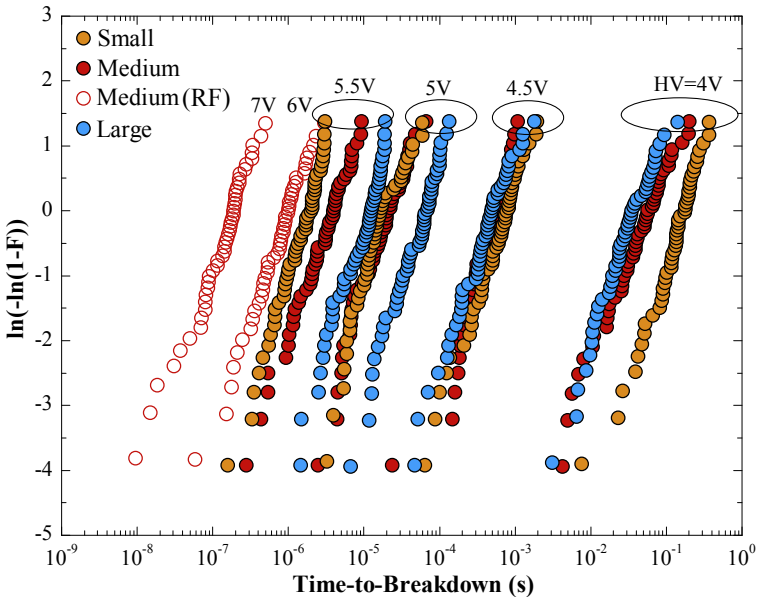


Figure 4.17: T_{BD} distributions for the Small, Medium and Large bitcells designed and fabricated in a logic 40-nm CMOS process.

Weibull slopes of the distributions are given in table 4.6

HV (V)	4	4.5	5	5.5	6	7
$\beta_{\mathbf{Small}}$	1.53	1.51	1.62	1.71	-	-
$\beta_{\mathbf{Medium}}$	1.26	1.66	1.56	1.39	1.38	1.35
$\beta_{\mathbf{Large}}$	1.46	1.46	1.64	1.65	-	-

Table 4.6: Weibull slopes of T_{BD} distributions for the Small, Medium and Large bitcells

Like in the previous examples, the slope β is similar for each distribution. The comparison of the three architectures is based on the T_{BD} distributions in a voltage range from 4 to 5.5V. In low voltage, the small bitcell exhibit a longer mean T_{BD} value than the large bitcell. This observation can be justified by the difference in capacitor area. Indeed, the area scaling law dictates that the larger the device area, the shorter the T_{BD} value. This trend was previously identified in section 4.3.

The behavior for HV=5.5V is opposite than in a low voltage range. For this condition, the small bitcell exhibits a shorter T_{BD} than the large bitcell. From this first observations, it can be seen that the dimensions of the antifuse bitcell impact strongly the time-to-breakdown.

The mean T_{BD} values were extracted from the distributions and plotted in figure 4.18 in order to identify the power-law models for each bitcell. The voltage drop across the series elements such as the drift transistor were taken into account to calculate the effective voltage across the capacitor.

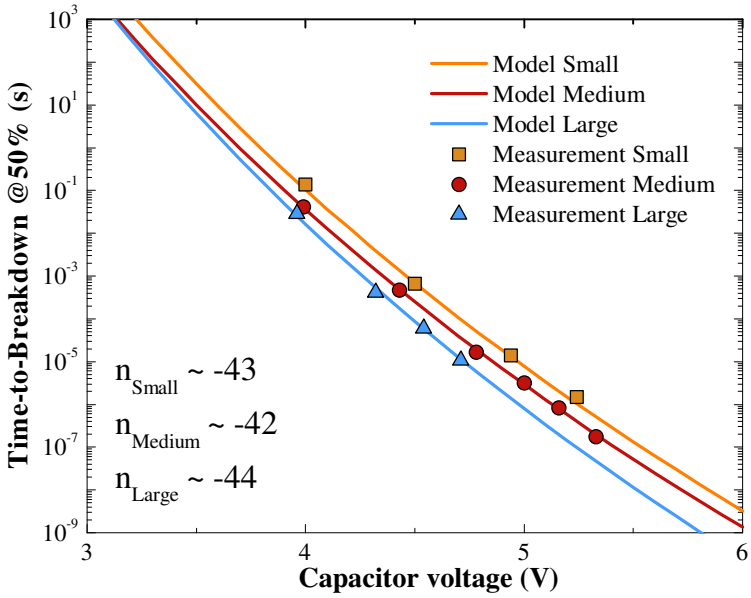


Figure 4.18: T_{BD} (V_{cap}) fitted against measurements for the Small, Medium and Large bitcells.

The modeling of the voltage-acceleration of T_{BD} is consistent with a power law for the small, medium and large bitcells. The acceleration factors differ slightly but remain in the range reported in literature, e.g. -43 ± 3. Even though this difference may lead to discrepancies if the models are extrapolated in a higher voltage range or for different bitcell dimensions, the precision should be sufficient to identify the main trends in the dependence of the time-to-breakdown in the programming voltage amplitude and device dimensions.

The area-scaling property is verified in this example as the larger the capacitor area, the shorter the time-to-breakdown value.

4.4.2.3 Results

The different parameters required to compute the capacitor voltage as a function of HV using equation (4.8) were identified in the previous steps. Consequently, V_{cap}

was computed in a HV range from 3 to 7V for each bitcell dimension (see table 4.5). Then, the time-to-breakdown was calculated using the power-law models identified in figure 4.18.
As a first result, T_{BD} is plotted versus V_{cap} and HV in figure 4.19.

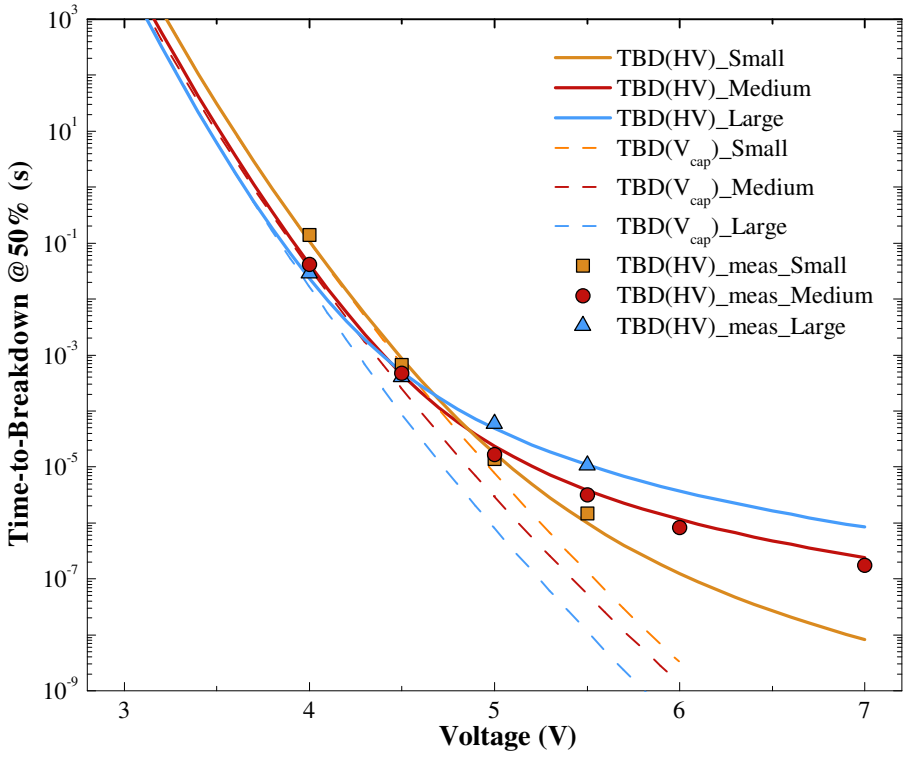


Figure 4.19: T_{BD} plotted versus V_{cap} and HV for the Small, Medium and Large bitcells. Measurements are correctly fitted with T_{BD} (HV) characteristics.

The characteristics T_{BD} (V_{cap}) were already plotted in figure 4.18 and are used for comparison with T_{BD} (HV). For the three bitcells, the two characteristics are coincident in a low voltage range. Then, T_{BD} (HV) deviates from the pure power-law model as the programming voltage amplitude is increased. It was shown that the voltage-acceleration of T_{BD} for single capacitors was equivalent while the influence of the capacitor area was reflected by a proportionality factor. A difference in the acceleration of T_{BD} (HV) curves is clearly emphasized in this graph. In fact, the time-to-breakdown is more accelerated for a small bitcell. For HV=7V, T_{BD} of the small bitcell is two orders of magnitude shorter than for the large bitcell whereas both devices exhibit a similar time-to-breakdown for HV=4.5V.
Since the trend on the voltage-acceleration of T_{BD} is in opposition to the area-scaling property, it is worth to compare the operating point of the three antifuse bitcells in the programming voltage range.
The modeling approach allows also the calculation of the wearout current using a

Fowler-Nordheim model. $I_{wearout}$ is plotted as a function of HV for the three bitcells in figure 4.20

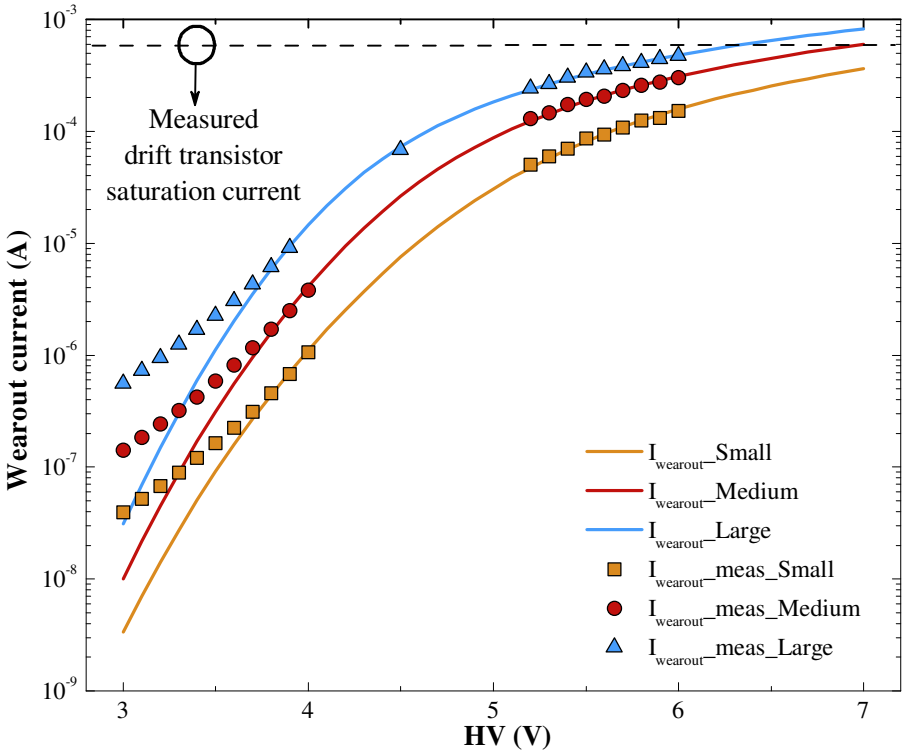


Figure 4.20: $I_{wearout}$ plotted versus HV for the Small, Medium and Large bitcells. Measurement are correctly fitted with the Fowler-Nordheim models.

The impact of the capacitor area is reflected by the amplitude of the wearout current for the three bitcells. The current of the small bitcell remains the lowest over the whole programming voltage range. Since the access transistor keeps the same dimensions, the effective voltage across the capacitor should be different. Due to a higher leakage current, the capacitor voltage of the large bitcell is lower than the voltage of the small bitcell.

The drift transistor is modeled by its on-state resistance as depicted in figure 4.16. This approximation leads to a limitation in the present methodology. Indeed, the saturation current of the drift transistor was characterized at 600μA which is not taken into account in the model. Moreover, the variation of the drain current with respect to the drain voltage is not purely linear around the pinch-off point. The calculation of the operating point if therefore not accurate. However, the approximation leads to an overestimated capacitor voltage and therefore a shorter time-to-breakdown. Considering the large bitcell, the wearout current should be limited by the drift transistor for a programming voltage higher than 6.2V. In this configuration, the capacitor voltage cannot be further increased as the difference in HV

amplitude is directly reported across the saturated drift transistor. Consequently, the time-to-breakdown should be nearly constant.

As previously introduced, the variation in voltage-acceleration between the three bitcells noticed on T_{BD} (HV) can be explained by studying the voltage operating point during the wearout phase, i.e. the voltage across the capacitor and the drift transistor for a given programming voltage amplitude.

4.4.2.4 Focus on the operating point

To compare the three bitcells, operating points are determined using Fowler-Nordheim models and I_{Drain} (V_{Drain}) drift transistor characteristics. An example is shown in figure 4.21.

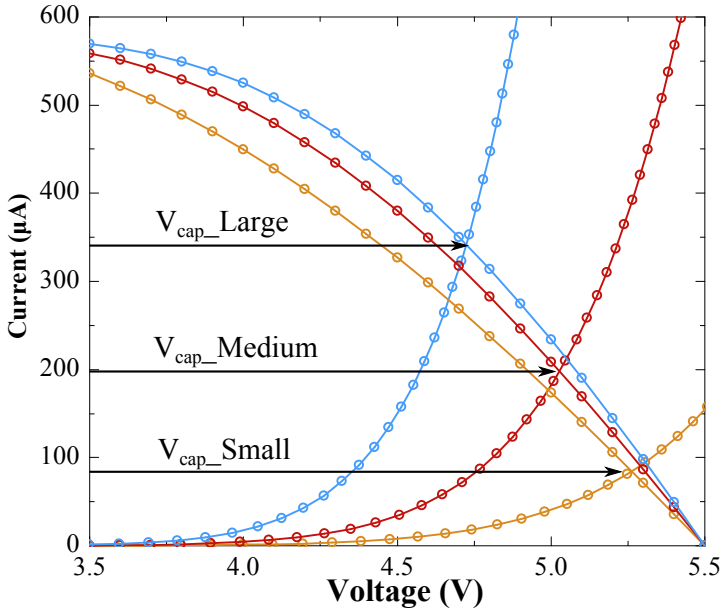
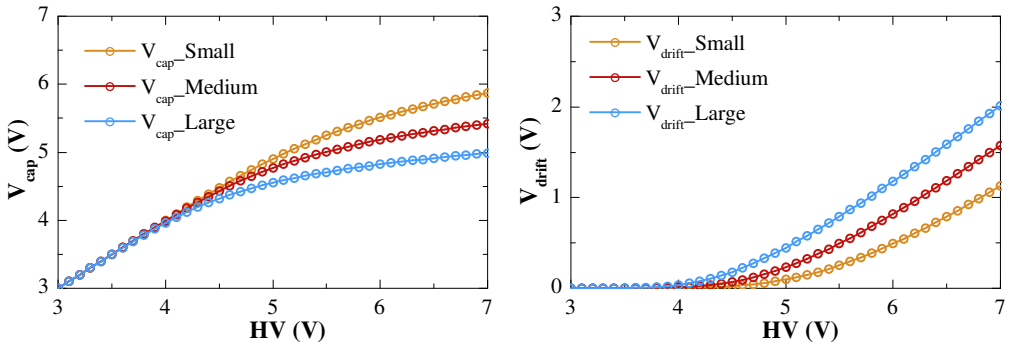


Figure 4.21: Voltage operating points for the small, medium and large bitcell. HV=5.5V.

Despite a slight difference in on-state resistance, it is seen that the capacitor voltage is mostly affected by the wearout current amplitude. Considering the small bitcell and HV=5.5V, $V_{cap} = 5.3V$ and $V_{Drift} = 200mV$ whereas for the large bitcell, $V_{cap} = 4.7V$ and $V_{Drift} = 800mV$. The capacitor voltage is therefore clearly affected by a large capacitor area. Since the time-to-breakdown is strongly accelerated by the capacitor voltage, the difference in capacitor area cannot overcome the voltage drop across the drift transistor caused by a higher wearout current. As a consequence, T_{BD} (HV) characteristic of the large bitcell is less accelerated in high voltage than the characteristic of the small bitcell.

The evolution of the operating point is further illustrated in figure 4.22. The capacitor and the drift transistor voltages are plotted versus the programming voltage amplitude.



(a) (b)

Figure 4.22: Capacitor voltage V_{cap} (a) and drift transistor voltage V_{drift} (b) versus HV for the small, medium and large bitcells.

The voltage drop across the drift transistor is negligible in a low voltage range. Hence, the capacitor voltage amplitude is approximately equivalent to the programming voltage. For HV higher than 4.5V, the significant wearout current leads to an increase in V_{Drift} and therefore the reduction of the capacitor voltage. The impact of the capacitor area is clearly seen as there is a difference of 1V between the capacitor voltage of the small and large bitcell for HV=7V.

4.4.3 Method of optimization

The high programming voltage necessary to program antifuse bitcells in a sufficiently short time leads to a high wearout current, a significant voltage drop across the access transistor and a high programming energy. Since the gate-oxide breakdown is strongly accelerated by the voltage applied across the dielectric, the optimization of the time-to-breakdown of an antifuse bitcell can be achieved by maximizing the capacitor voltage or minimizing the voltage drop across the access transistor.

A significant impact of the capacitor area was emphasized in the previous study. Practical measurements and an associated modeling methodology revealed that a bitcell featuring a small capacitor exhibits a shorter time-to-breakdown than a large capacitor in a high voltage range. The dimensions of the antifuse bitcells such as the capacitor area and the channel of the drift transistor have a prime role in the resulting time-to-breakdown.

4.4.3.1 Methodology and algorithm

The modeling approach presented in figure 4.16 allows the calculation of the time-to-breakdown according to the dimensions of the antifuse bitcell and the programming voltage amplitude.

The optimization consists in calculating the required programming voltage amplitude according to a targeted time-to-breakdown, the capacitor area and the on-state resistance of the access transistor. Thus, the impact of the bitcell dimensions on HV is studied in order to point out an optimum in terms of programming voltage amplitude and bitcell area. An algorithm is depicted in figure 4.24.

Input parameters. The wearout current is modeled as Fowler-Nordheim-like and the voltage-acceleration of the time-to-breakdown by a power law. Different parameters are identified from measurements to align the models with the CMOS technology in which the antifuse bitcells are fabricated. The wearout current model was identified using the measurement shown in figure 4.20.

A single power-law model was identified for the three bitcells although the acceleration factors were slightly different as shown in figure 4.18. Consequently, an average acceleration factor n of -43 is chosen with the corresponding proportionality factor α. Thus, the power-law model can be projected for various capacitor area. Resulting $T_{BD}\,(V_{cap})$ models are plotted in figure 4.23 for the small, medium and large bitcells.

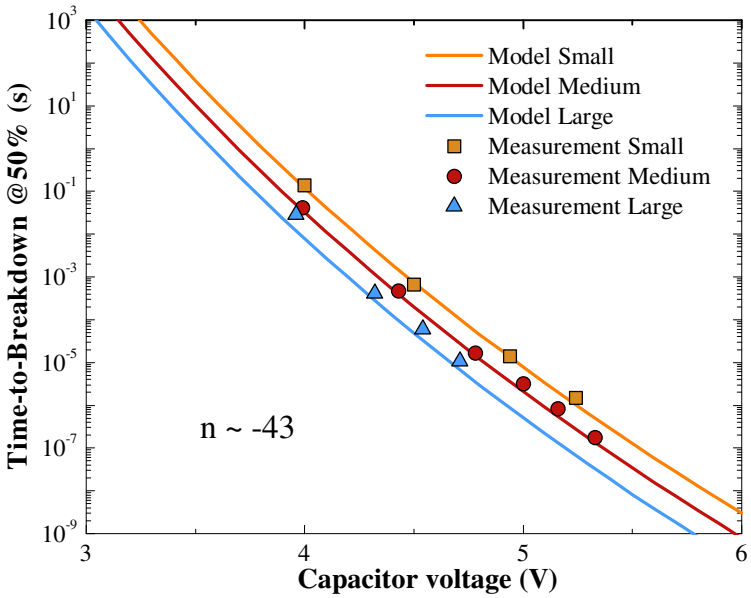


Figure 4.23: $T_{BD}\,(V_{cap})$ power-law models calculated for the small, medium and large bitcells fabricated in a logic 40-nm CMOS process.

The power-law models are correctly projected from a capacitor area to another. The fit against measurements is less accurate than the models shown in figure 4.18 which were identified for each device. However, the precision is sufficient for the purpose of the methodology.

Targeted T_{BD}. As introduced previously, this model aims at calculating the programming voltage amplitude required to achieve the gate-oxide breakdown of an antifuse bitcell in a given time. In the present example, T_{BD} is set to 100ns.

A_{cap} range. Calculations are performed for different capacitor area values. A range from $0.1\mu m^2$ to $3\mu m^2$ is chosen which corresponds to a capacitor width W_{cap} range from 0.4μm to 12.2μm and a constant length of 0.3μm.

R_{on} range. The impact of the on-state resistance of the access transistor is also studied. A range from 500Ω to 5000Ω is chosen. The corresponding transistor width W_{drift} for a constant length of 0.3μm is from 1μm to 10.5μm.

V_{cap} calculation. The capacitor voltage is calculated as a function of the targeted time-to-breakdown, the capacitor area and the parameters of the power law are identified from reference measurements.

Equation solver. Finally, the equation of HV is computed using an equation solver for the different capacitor area and on-state resistance values defined previously. V_{cap} was calculated in a previous step and the wearout current is modeled using a Fowler-Nordheim approach. The parameters C and D were identified from previous measurements.

Model output. As a result, the model yields the programming voltage amplitude HV required to reach the targeted time-to-breakdown (100ns) in a capacitor and drift transistor width range.

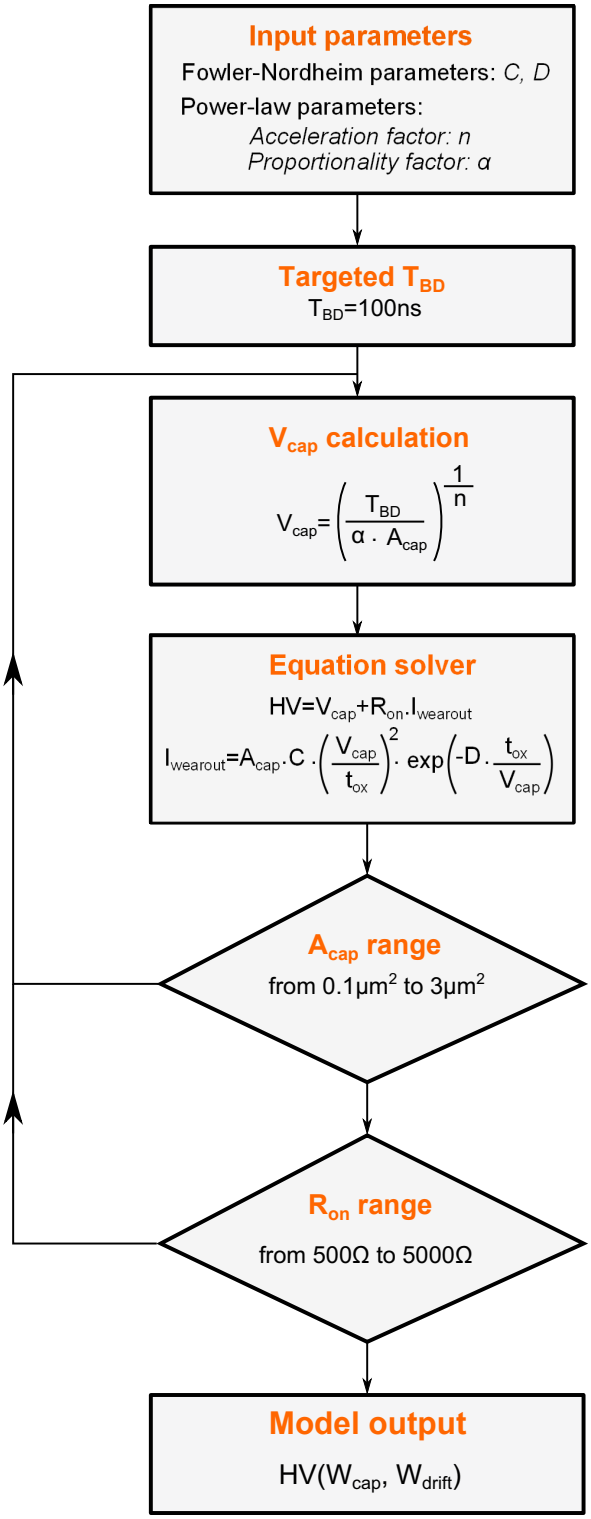


Figure 4.24: Algorithm for the calculation programming voltage amplitude according to a targeted T_{BD}, the capacitor width W_{cap} and the drift transistor width W_{drift}.

4.4.3.2 Results

The programming voltage amplitude computed using the algorithm presented previous is plotted as a function of both the capacitor area and the transistor width in figure 4.25.

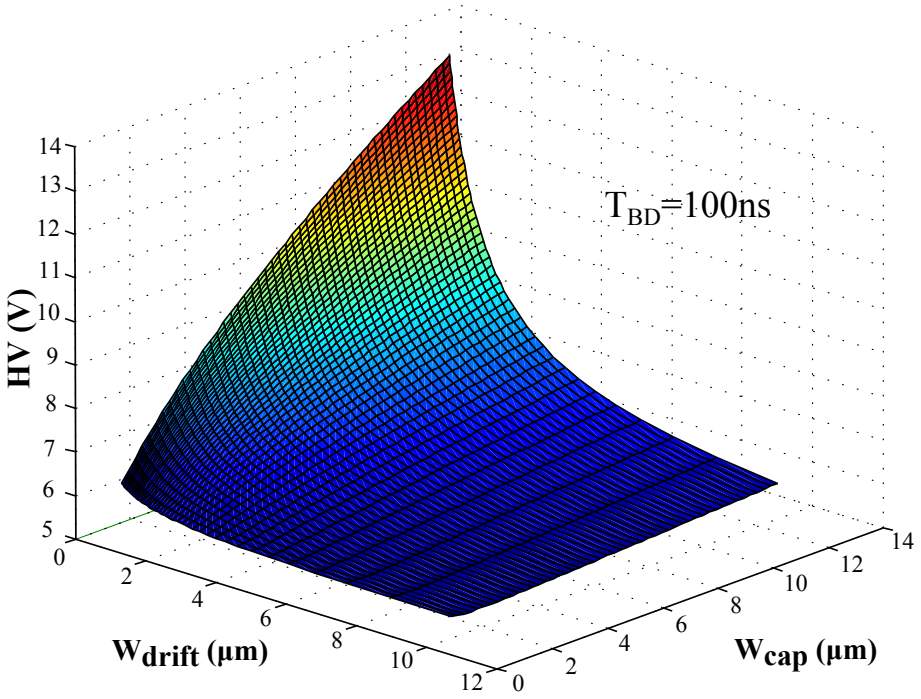


Figure 4.25: HV programming voltage versus the capacitor width W_{cap} and the drift transistor width W_{drift} for T_{BD} = 100ns computed using a TDDB model.

The programming voltage amplitude HV is strongly impacted by the dimensions of the antifuse bitcells. Indeed, it varies from 5.6 to 13V. The minimum HV is obtained for the maximum drift transistor width and approximately the minimum capacitor dimension. This result is explained by the following equation of the programming voltage amplitude used in the algorithm depicted in figure 4.24.

$$HV = V_{cap} + R_{on} \cdot I_{wearout} \tag{4.11}$$

The short targeted T_{BD} value leads to a high capacitor voltage V_{cap} and therefore high wearout current. Consequently, R_{on} has a direct impact on the resulting HV amplitude as the lower the on-state resistance, the lower the voltage drop across the access transistor and thereby the higher V_{cap} for a given HV. The minimum programming voltage amplitude is obtained for a low R_{on}, i.e. a large transistor, and a low $I_{wearout}$, i.e. a small capacitor. Although a large capacitor exhibits a shorter T_{BD} as shown in figure 4.23, the voltage-acceleration term (n=-43) is more

significant than the proportionality factor. This is the reason why the contribution of the wearout current is the dominating factor.
This statement is further illustrated in figure 4.26. HV (W_{cap}) is plotted for a maximum W_{drift} while HV (W_{drift}) is plotted for a minimum W_{cap}. T_{BD} is set to 10ns, 100ns and 1μs.

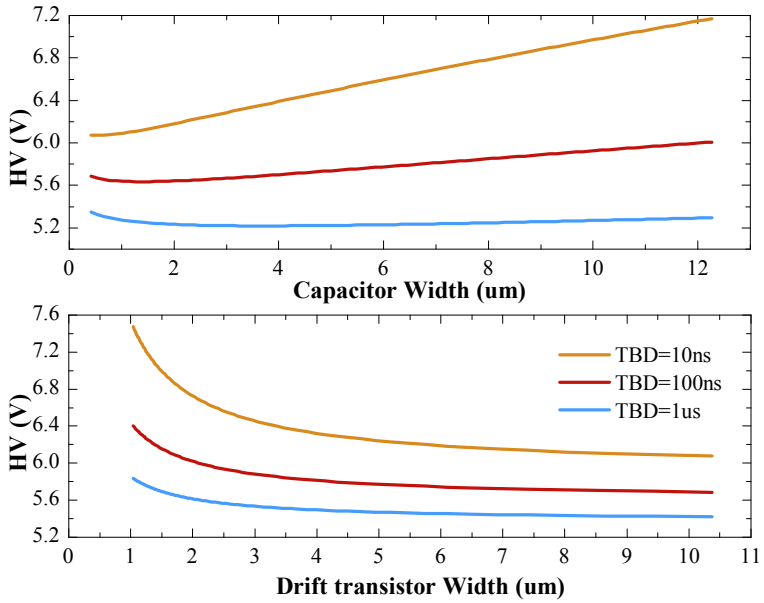


Figure 4.26: HV programming voltage calculated using the TDDB model and plotted versus W_{cap} and W_{drift} for T_{BD} =10ns, 100ns and 1μs and for constant W_{drift} (top) and W_{cap} (bottom).

In the three cases, the minimum HV is obtained for a small capacitor width. However, W_{cap} is slightly higher for $T_{BD} = 1\mu$s. Since the capacitor voltage is lower, the term $R_{on} \cdot I_{wearout}$ is no longer significant. The proportionality factor is dominating leading to higher programming voltage for a smaller capacitor area.
For any T_{BD}, the minimum HV value is obtained for the largest W_{drift}, i.e. the lower R_{on}. However, there are two limitations in using a large drift transistor. First, there is an obvious penalty in density. Second, a large channel leads to a high post-breakdown current. Even though a sufficient post-breakdown current is required to onset a satisfying read current, it has a direct impact on the power consumption and on the area occupied by the peripheral circuitry. Since the gain in HV amplitude along W_{drift} is not significant, the definition of this dimension must be counterbalanced by the impact on the post-breakdown current.

4.4.4 Conclusion

A Time-Dependent Dielectric Breakdown model dedicated to antifuse bitcells was presented in this section. Due to the high voltage range, the wearout current dam-

aging the dielectric leads to a significant voltage drop across the access transistor. This is the reason why the accurate determination of the capacitor voltage is essential to achieve a relevant TDDB model for antifuse bitcells.
The present TDDB model aims at calculating the effective capacitor voltage as a function of the programming voltage amplitude, the capacitor area and the dimensions of the access transistor. As a result, it was shown that an antifuse bitcell featuring a small capacitor exhibits a shorter time-to-breakdown than a large capacitor in a high voltage range. This trend is confirmed by the TDDB model and emphasized by the difference in the effective capacitor voltage amplitude.
According to the previous observation, an optimization of the antifuse bitcell design consists in the maximization of the capacitor voltage. In other words, the voltage drops across the series elements such as the access transistor must be minimized. An optimization methodology was developed which aims at computing the programming voltage amplitude as a function of a targeted time-to-breakdown, the capacitor area and the drift transistor dimensions. The benefit of a small capacitor was further emphasized by this model while the access transistor has a second-order impact.
With the increasing demand in dense antifuse memories, the reduction of the capacitor area is obviously valuable. Moreover, it has a positive impact in the peripheral circuitry as the programming voltage amplitude is also reduced.
Finally, the impact on the optimization of the programming energy is discussed in conclusion of the chapter.

4.5 Cascode antifuse bitcell

4.5.1 Architecture and performance

A cascode antifuse bitcell comprises a MOS capacitor connected in series with two regular MOS transistors. A schematic and cross-section views are depicted in figure 4.27.

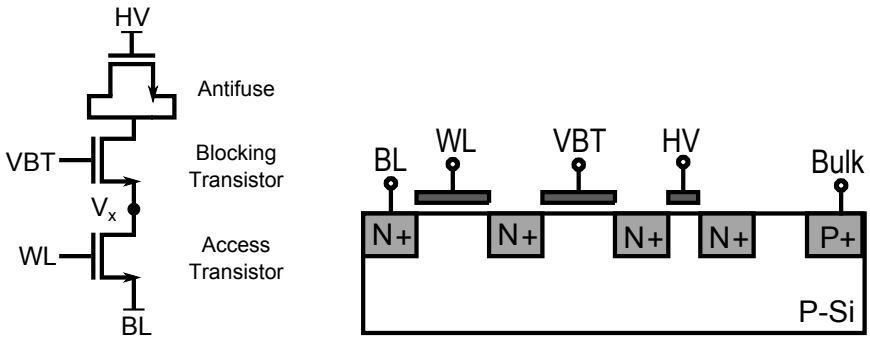


Figure 4.27: Schematic and cross section of a cascode antifuse bitcell. VBT=4.25V, WL=1.8V and BL=0V in the programming mode.

The present cascode antifuse bitcell features only N-type devices. Unlike the drift bitcell, a Nwell is not necessary thereby allowing a denser structure. As presented in chapter 2, section 2.3.2, the high voltage applied to the drain of the access device is divided across the two transistors. Therefore, a proper dimensioning and the blocking voltage (VBT) amplitude insure the reliability of the bitcell.
The dimensions of a cascode antifuse bitcell designed and fabricated in a logic 40-nm CMOS process are given in table 4.7.

$\mathbf{A_{bitcell}}$	$\mathbf{A_{cap}}$	$\mathbf{W_{BT}}$	$\mathbf{W_{AT}}$
$4\mu m^2$	$0.041\mu m^2$	$1.4\mu m$	$1.4\mu m$

Table 4.7: Dimensions of a cascode antifuse bitcell designed and fabricated in a logic 40-nm CMOS process.

The antifuse capacitor is a thin-oxide transistor featuring the minimum length allowed in this technology. Hence, the capacitor area is significantly smaller than the drift bitcell studied in section 4.3.3 ($10\mu m^2$). The thick-oxide cascode transistors have a longer length and are dimensioned for an appropriate saturation current.
Time-to breakdown measurements were performed in cascode antifuse bitcells using a RF experimental setup. T_{BD} distributions are plotted in figure 4.28 for programming voltage amplitudes of 5.5, 6 and 6.6V respectively. In addition, the distribution of the 10-μm^2 drift bitcell programmed under 7V is plotted for comparison. Table 4.8 gives the Weibull slopes β_{cascode}.

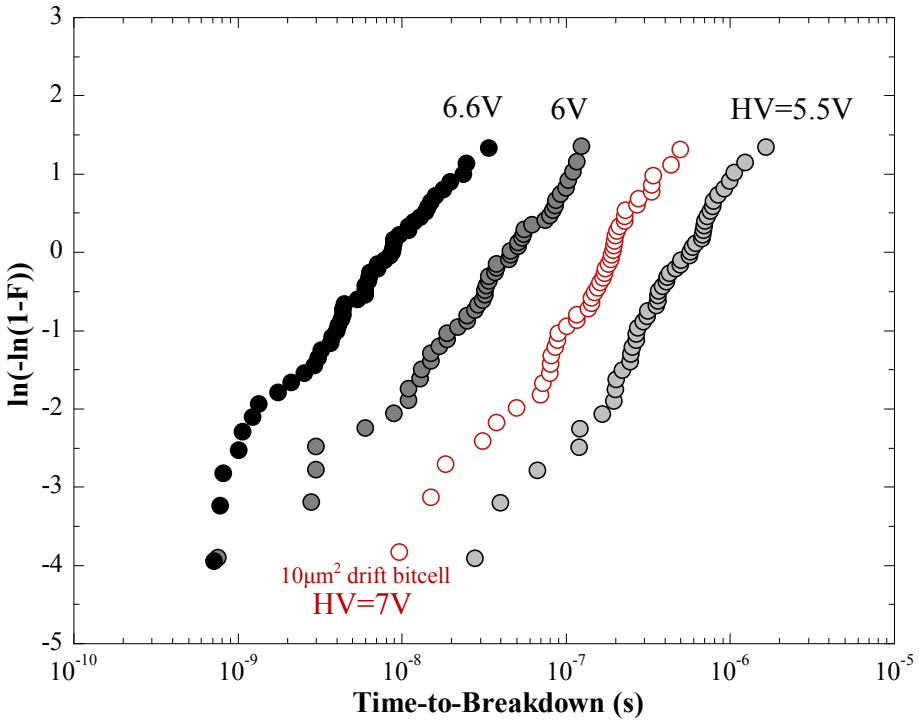


Figure 4.28: T_{BD} distribution for 4-μm^2 cascode antifuse bitcells designed and fabricated in a logic 40-nm CMOS process. An additional distribution of a 10-μm^2 drift bitcell is plotted for comparison purpose.

HV (V)	**5.5**	**6**	**6.6**
$\beta_{\mathbf{cascode}}$	1.4	1.1	1.2

Table 4.8: Weibull slopes of T_{BD} distributions for a cascode antifuse bitcell fabricated in a logic 28-nm CMOS process.

The cascode antifuse bitcell exhibits very good performances in term of time-to-breakdown. The comparison of this distribution with the measurements performed on a 10-μm^2 drift antifuse bitcell under the product programming conditions (HV=7V) shows a difference of one order of magnitude on the mean T_{BD} value. This difference in performance between the cascode and the drift antifuse bitcell can be explained using the modeling approach presented in section 4.4. The capacitor area of the cascode bitcell is more than ten times smaller than in the drift bitcell ($A_{cascode} = 0.041\mu m^2$ and $A_{drift\_10\mu m^2} = 0.58\mu m^2$). Consequently, the wearout current amplitude is lower. Since the access transistor is composed of two regular thick-oxide MOS transistor, the on-state resistance is also lower ($R_{on\_cascode} = 2k\Omega$, $R_{on\_drift\_10\mu m^2} = 2.6k\Omega$). The voltage drop across the access device $R_{on} \cdot I_{wearout}$ is therefore minimized in the cascode architecture. This is the reason why performance in time-to-breakdown is improved.

4.5.2 High-K cascode antifuse bitcell

A major innovation is involved in the 32-nm/28-nm technology node because the SiO_2 gate-oxide is replaced by a high-K material. In was presented in chapter 3, section 3.3 that the results from TDDB characterization slightly differ from the common models and observations proven on fully-silicon devices. Nevertheless, the breakdown of high-K dielectric materials is exploitable for antifuse memories.
For providing insight into the performance of these bitcells, time-to-breakdown and wearout current measurements where performed on cascode antifuse bitcells designed and fabricated in a logic 28-nm CMOS process.
The schematic and the cross-section are the same as shown in 40-nm CMOS in figure 4.27. The dimensions of the antifuse bitcells characterized in this study are given in table 4.9. The bitcells are denoted as A or B according to the capacitor area.

	$\mathbf{A_{cap}}$	$\mathbf{W_{BT}}$	$\mathbf{W_{AT}}$
A	$0.027\mu m^2$	$0.9\mu m$	$0.9\mu m$
B	$0.036\mu m^2$	$0.9\mu m$	$0.9\mu m$

Table 4.9: Dimensions of a cascode antifuse bitcell as designed and fabricated in a logic 28-nm CMOS process.

4.5.2.1 Wearout current measurements

The replacement of the silicon dioxide by a high-K material has obviously an impact on the wearout current. Consequently, wearout current measurements were performed on cascode antifuse bitcells using a DC and transient experimental setup and are plotted in figure 4.29. The wearout current characteristics of the 10-μm^2 drift and the cascode bitcells fabricated in a logic 40-nm CMOS process and presented previously are also plotted for comparison.

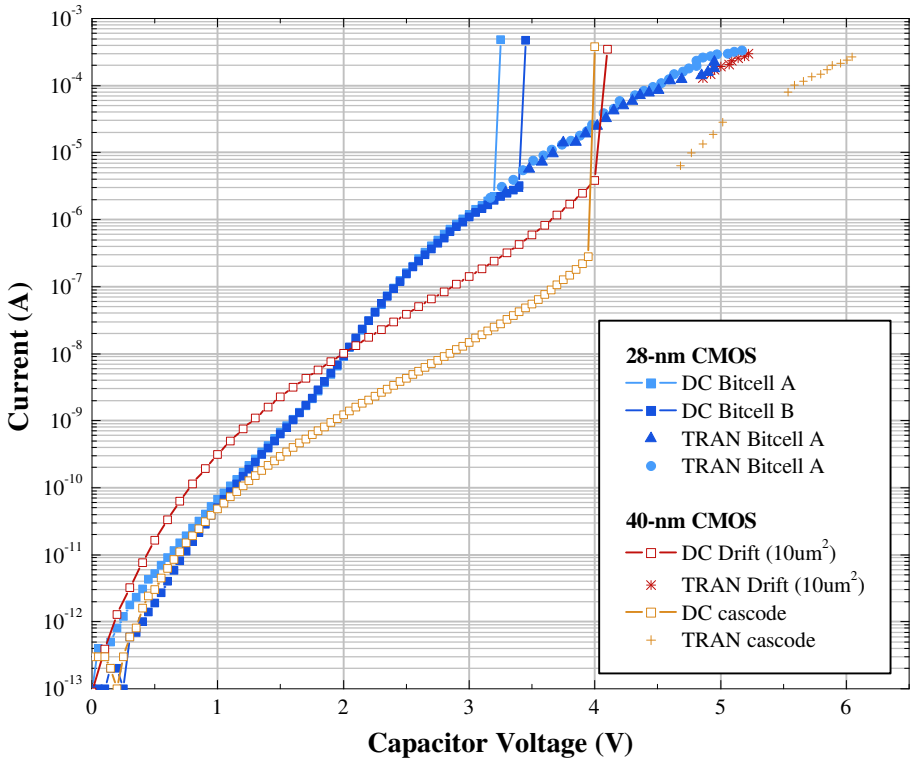


Figure 4.29: DC and transient wearout current measurements performed on cascode antifuse bitcells (A and B in table 4.9) designed and fabricated in a logic 28-nm high-K/metal-gate CMOS process. Measurements performed on drift and cascode bitcells fabricated in 40-nm CMOS are plotted for comparison purpose.

Due to the small difference in capacitor area, the wearout current amplitude of the bitcells A and B are somewhat similar. The breakdown voltage of the high-K material is significantly lower than the fully-silicon capacitor. The capacitor seems therefore more fragile. The robustness of the dielectric is investigated further in the document.

The amplitude of the wearout current is greatly higher in a high voltage range than for the fully-silicon bitcells whereas the capacitor area of the cascode bitcells A and B are the smallest. Finally, the 28-nm cascode bitcells and the 40-nm drift bitcell exhibit a similar wearout current for a capacitor voltage slightly higher than 5V. Since the 28-nm capacitor are is more than ten times smaller than the capacitor area of the drift antifuse bitcell, it can be concluded that the high-K stack is drastically leakier than the conventional silicon dioxide in the programming voltage range of antifuse bitcells. The small dimension of the antifuse capacitor in the 40-nm cascode bitcell leads to a wearout current lower than in the drift and the 28-nm cascode bitcells. Indeed, a leakage current of 300μA is reached for a voltage of 6V whereas a similar current is measured for a capacitor voltage slightly higher than 5V on the

other devices.

The shape of the wearout current is also different in 28-nm cascode bitcells. Two modes can be distinguished for a capacitor voltage lower or higher than 2V. Since the dielectric stack is composed of a SiO_2 interface layer beneath the high-K material, the leakage current depends on the current transport process occurring in both materials. It was shown that carriers can cross the interface layer due to a direct tunnel mechanism whereas a trap-assisted transport can occur through the high-K material [110, 111].

4.5.2.2 Time-to-breakdown measurements

Time-to-breakdown measurements were performed using a RF setup on the bitcells A and B. Distributions on a set of 50 bitcells are plotted in figure 4.30. Table 4.10 gives the Weibull slopes.

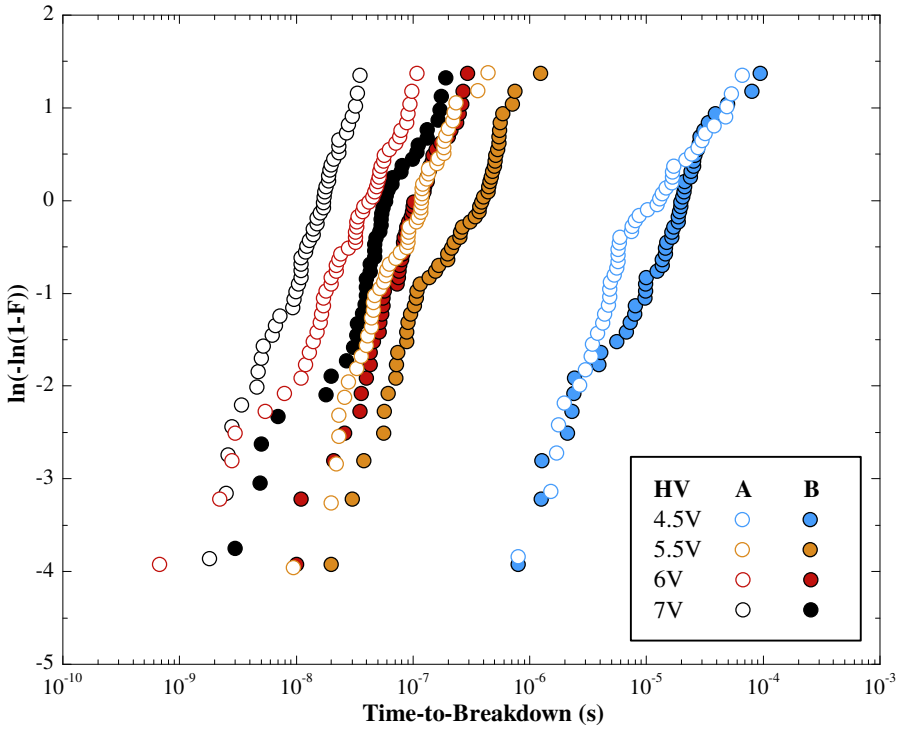


Figure 4.30: T_{BD} distributions of cascode antifuse bitcells (A and B in table 4.9) designed and fabricated in a logic 28-nm high-K/metal-gate CMOS process.

HV (V)	**4.5**	**5.5**	**6**	**7**
$\beta_{\mathbf{A}}$	1.1	1.4	1.1	1.6
$\beta_{\mathbf{B}}$	1.1	1.2	1.5	1.5

Table 4.10: Weibull slopes of T_{BD} distributions for a cascode antifuse bitcell fabricated in a logic 40-nm CMOS process.

The mean T_{BD} values of the distributions for HV=4.5V are slightly different. However, the time-to-breakdown appears more accelerated for the bitcell A which features a smaller capacitor area than the bitcell B. For HV=7V, the mean T_{BD} value of bitcell A is 14ns while bitcell B gives 48ns.

The benefit of a small capacitor in the reduction of the time-to-breakdown seems also relevant for the antifuse bitcells featuring a high-K dielectric. Since the high-K material appears leaky, the optimization of the voltage operating point is essential. Indeed, the voltage-acceleration of the time-to-breakdown is severely affected in high voltage. Considering the distributions for HV=6 and 7V for both bitcells, it is seen that the mean T_{BD} value is slightly shorter. The wearout current amplitude measured in this high voltage range reaches nearly the saturation current of the cascode transistor. Hence, the capacitor voltage cannot be significantly increased by the programming voltage amplitude.

4.5.2.3 Discussion

The cascode architecture shows very promising properties. The combination of a small capacitor and a cascode transistor with a low on-state resistance allows the maximization of the capacitor voltage. Moreover, a cascode antifuse bitcell is denser than a drift bitcell because it only comprises regular NMOS transistors which do not feature a bulky and resistive Nwell. On the other hand, an additional voltage is required to drive the blocking transistor which is higher than the conventional supply voltage available in SoC. The voltage generator is therefore needed, thereby leading to a more complex memory system.

The integration of a high-K material in the CMOS process brings a major change in the structure of the antifuse capacitor. Even though the dielectric stack has become thicker, characterizations emphasized that the high-K dielectric is less robust and leakier than the silicon dioxide. Furthermore the wearout current and time-to-breakdown models are different and are currently reported in literature. Nevertheless, the breakdown of high-K dielectric is strongly accelerated by the stress voltage. The optimization of the bitcell dimensions consisting in maximizing the capacitor voltage remains therefore consistent.

The methods of characterization presented in this Ph.D. work are applicable to investigate the breakdown physics of high-K dielectric in the programming voltage range of antifuse memories.

4.6 Conclusion

The characterization and the modeling of the time-to-breakdown of antifuse bitcells is particular due to the high programming voltage range and the resulting short T_{BD}. Dedicated experimental setup were put in place in order to attain a bandwidth allowing minimum T_{BD} measurements of 1ns. This can be achieved down to 1μs using a series resistor or down to 1ns using a RF bias-Tee. Time-Dependent Dielectric breakdown (TDDB) characterizations were performed for a maximum programming of 7V which is the amplitude used in industrial product designed in logic 45-nm and 40-nm CMOS processes.

The measurements of the wearout phase from the application of the programming voltage to the breakdown event highlighted a significant leakage current. Hence, it turns out that a dramatic voltage drop is lost across the access device. Since the breakdown mechanisms are accelerated by the voltage applied across the dielectric material, it is worth to determine accurately the capacitor voltage in order to identify a relevant TDDB model.

The wearout current was successfully modeled using a Fowler-Nordheim law. Since the defect density in ultrathin oxide is very low, a conduction mechanism limited by the barrier heights of the structure appeared more relevant than trap-assisted mechanisms. Moreover, the high programming voltage leads to a severe band-bending which is in favor of a Fowler-Nordheim conduction mode rather than Direct tunnel. Fowler-Nordheim models were correctly fitted with measurements. Furthermore, it was shown that a model identified on a reference device can be extrapolated to different capacitor area. This property is very useful in order to investigate the influence of the capacitor area on the voltage operating point of the antifuse bitcell.

The modeling of the time-to-breakdown is a major objective. As mentioned previously, the voltage applied across the capacitor must be determined in order to identified proper models from measurements. First, TDDB characterizations were performed on a set of 50 bitcells for each programming voltage amplitude. Thus, T_{BD} distributions were plotted on a Weibull scale. The slopes were consistent with the common observations on ultrathin-oxide capacitors in the high programming voltage range of antifuse memories. Second, the mean T_{BD} values were plotted as a function of the capacitor voltage which was determined using prior wearout current measurements. Then, a power law was identified with an acceleration factor similar

to the characteristic reported in literature. The area-scaling property was also verified and the power-law model identified from a reference device could be projected for different capacitor areas.

Since the wearout current and the time-to-breakdown were modeled using a Fowler-Nordheim and a power law respectively, the impact of the antifuse bitcell dimensions on the wearout phase was investigated. An analytical model was built up which aims at calculating the capacitor voltage as a function of the capacitor area, the on-state resistance of the access transistor and the programming voltage amplitude. The pertinence of this approach was verified experimentally on drift antifuse bitcells featuring different capacitor areas and the same access transistor dimensions. TDDB characterizations has shown that the shortest time-to breakdown was obtained for a bitcell architecture with the smallest capacitor. This experimental result was confirmed by the analytical model as the capacitor voltage was much higher for a small capacitor. Indeed, the wearout current is lower, so is the voltage drop across the access device. This trend is valuable to reduce the size of the bitcell. Another modeling approach has shown that for a targeted T_{BD} value, the usage of a small capacitor leads to a low programming voltage amplitude. As a consequence, an optimum bitcell allows also the reduction circuit area occupied by the charge-pump HV generator.

The programming mechanisms of antifuse bitcell fabricated in a logic 28-nm high-K/metal-gate CMOS process were also investigated. In addition to this particularity, a cascode bitcell architecture is used. TDDB characterizations were performed in a cascode bitcell fabricated in a 40-nm CMOS process prior to 28-nm bitcells. The cascode bitcell features a smaller capacitor than a drift bitcell because the gate is dimensioned with a minimim length, e.g. 40nm. The measurements have confirmed the benefit of the small capacitor as well as the lower on-state resistance of cascode transistor. TDDB characterizations performed on 28-nm bitcells revealed that the dielectric stack is less robust and leakier than the silicon dioxide. Due to the high wearout current, the performances are between the cascode and the drift 40-nm bitcells. The measurements have confirmed that the modeling of the wearout current and the time-to-breakdown are more complex. However, the characterization methodologies put in place in this Ph.D. work can be used to further investigate the breakdown physics of high-K stack. For now, it is obvious that breakdown physics is accelerated by the capacitor voltage. The operating voltage must be therefore optimized for 28-nm bitcells. The accurate TDDB modeling requires extra work that are not discussed in this Ph.D. thesis.

The numerous TDDB characterizations performed on a variety of antifuse bitcells have shown that the wearout phase is very short compared to the programming

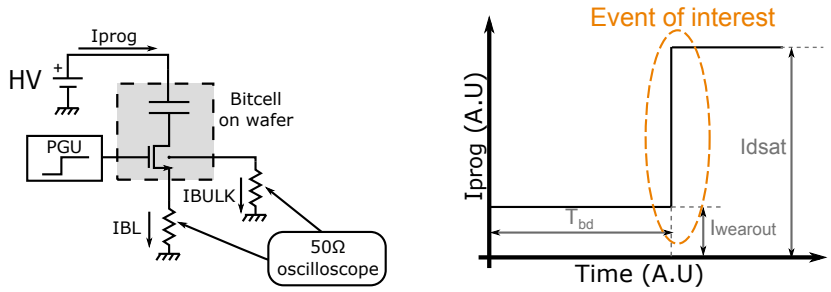
time reported in literature. Considering a programming time of 10μs, the time-to-breakdown is in a range between 2ns and 35ns for a wearout current amplitude of 300μA. The maximum charge during the wearout phase is therefore 10pC. However, the charge in the post-breakdown phase, assuming a saturation current of 350μA is 3.5nC. This simple calculation demonstrates that most of the charge and therefore, most of the programming energy is consumed in the post-breakdown phase. The study and the modeling of the wearout phase was finally helpful to dimension the antifuse bitcell in order to reduce the time-to-breakdown and the programming voltage amplitude. The optimization of the programming energy is related to the post-breakdown phase and is discussed in details in chapter 5.

CHAPTER 5

Side effect: Bulk current overshoot

The author thanks Hervé Morel, Dominique Planson and Dominique Tournier from Ampere Lab for significant and helpful support in the understanding of the here presented phenomenon.

Time-to-breakdown measurements are performed on antifuse bitcells by applying a constant voltage stress across the capacitor and by probing the current flowing through the dielectric. The bitline and the bulk current are measured using the setup depicted in figure 5.1(a). A schematic waveform of the programming current is proposed figure 5.1(b).



(a) (b)

Figure 5.1: Experimental setup used to measure the bitline and the bulk current (a). Schematic of the evolution of the capacitor gate current (b).

As presented in chapter 4, a tunneling current is measured in the bitline node (I_{BL}) whereas the bulk current (I_{Bulk}) is negligible during the wearout phase. At the breakdown event, the bitline current rises up to the saturation level of the access

transistor while the bulk reaches an amplitude higher than the bitline current. The cause of this high bulk current overshoot is discussed in this chapter.

The sum of the bitline and bulk currents gives a value larger than the saturation current of the access transistor. The schematic approach of the programming current depicted in figure 5.1(b) is then erroneous. A practical example of measurement performed on an antifuse bitcell designed and fabricated in a logic 45-nm CMOS process is shown in figure 5.2.

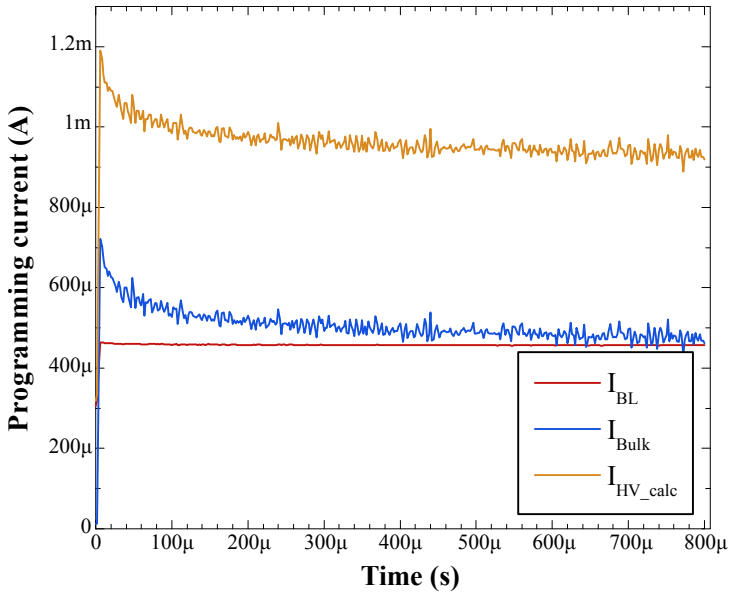


Figure 5.2: Measurements of the bitline and bulk currents on a drift antifuse bitcell fabricated in a logic 45-nm CMOS process during a constant voltage stress (HV=6V). The HV current is calculated as $I_{HV\_calc} = I_{BL} + I_{Bulk}$ since there is no significant current detected in the rear contact of the wafer to the prober chuck connected to GND.

For HV=6V, the time-to-breakdown is about 1μs whereas the time scale of the bulk current overshoot is longer than hundreds of microseconds. The bitline current is, as expected, limited by the drift transistor ($I_{dsat} = 450\mu A$). The bulk current has a particular shape. A very steep increase up to 720μA is seen at breakdown. Then, the bulk current decays for 300μs down to roughly I_{dsat}.

Before discussing the root cause of the observed bulk current overshoot, it is worth to point out its impact on the antifuse memory system.

- **Power consumption:** The maximum programming current is assumed to be limited by the drift transistor, e.g. $I_{dsat} = 450\mu A$. However, the bulk current overshoot leads to a peak amplitude of 1.2mA and an average value of 1mA for 800μs. The power consumption is therefore severely impacted.

- **HV source specifications:** The output current specification of the charge-pump circuit dedicated to programming is specified according to the saturation current of the access transistor. Consequently the high voltage generator would be undersized and unable to deliver the sufficiently high output current and the required high voltage amplitude. Therefore, the programming time could be impacted.

- **Read current:** The performance of an antifuse memory in terms of access time and yield is determined by the read current amplitude. In fact, the read current amplitude depends on the breakdown spot impedance which is set by the programming conditions, e.g. the programming current amplitude. The read current may be impacted by the bulk current overshoot. This latter point is discussed in chapter 5.

According to the few examples presented above, it appears obvious to study the cause of the bulk current overshoot occurring during the programming operation.
The facts are described in section 5.1.
Results from electrical characterizations are detailed in section 5.2. The impact of various programming parameters are studied such as the programming and wordline voltages, the temperature and the bulk biasing.
An assumption of the root cause of the bulk current is investigated in section 5.3.
Conclusions are drawn in section 5.4

5.1 Facts

The different antifuse devices in which the phenomenon has been noticed are presented in this section.

5.1.1 Antifuse devices

The study is focused on drift antifuse bitcells. As introduced previously, the bulk current overshoot was observed during the post-breakdown phase. Consequently, it is interesting to describe accurately the bitcell architecture and the possible carrier flow from the gate of the antifuse capacitor to the bulk node.
A schematic and a cross-section of a drift antifuse bitcell are depicted in figure 5.3.

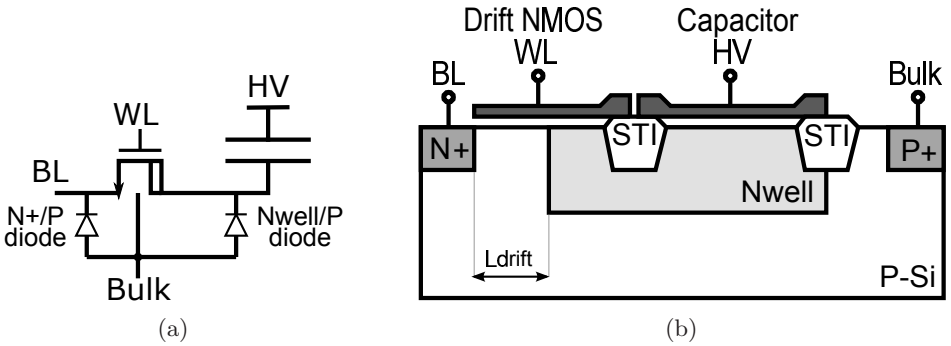


Figure 5.3: Schematic and cross section of a drift antifuse bitcell including body diodes.

Table 5.1 gives the dimensions of antifuse bitcells designed and fabricated in logic 55-nm, 45-nm and 40-nm CMOS processes. The devices were used to investigate the underlying physical mechanism leading to the bulk current overshoot.

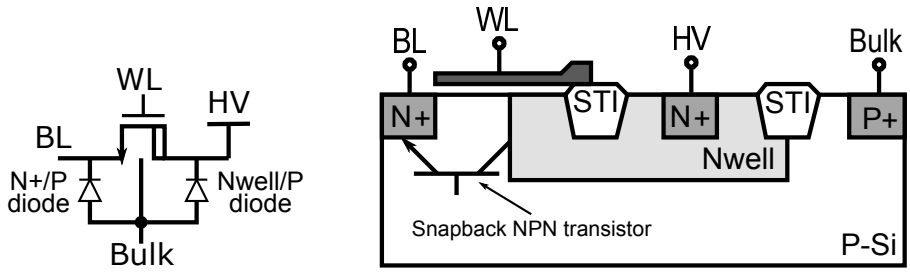
CMOS	$A_{bitcell}$	A_{cap}	W_{drift}	L_{drift}
55nm	$13\mu m^2$	$0.80\mu m^2$	$1.98\mu m$	$0.76\mu m$
45nm	$16\mu m^2$	$0.97\mu m^2$	$2.2\mu m$	$0.85\mu m$
45nm	$5\mu m^2$	$0.13\mu m^2$	$0.5\mu m$	$0.25\mu m$
40nm	$10\mu m^2$	$0.58\mu m^2$	$2\mu m$	$0.3\mu m$

Table 5.1: Dimensions of antifuse bitcells as designed and fabricated in logic 55-nm, 45-nm and 40-nm CMOS processes.

There are three possible paths for the bulk current. The diodes at Nwell/P-substrate and N+/P-substrate junctions are reverse biased. Considering the amplitude of the bulk current, such a leakage current is unlikely from a reverse-biased diode unless a punch-through or a avalanche mechanism occurs. A hot carrier current from the drift transistor is also a possible cause of the bulk current. Assumptions are discussed in the following subsections.

5.1.1.1 Single drift

Characterizations were performed on single drift transistors in order to identify whether the bulk current overshoot occurs. The gate of the antifuse capacitor is replaced by a contact on the Nwell. Consequently the high voltage is directly applied to the drain of the drift transistor. A schematic and a cross-section of a drift transistor are depicted in figure 5.4.



(a) (b)

Figure 5.4: Schematic (a) and cross section (b) of a single drift transistor including body diodes.

A bulk current can occur in a MOS transistor due to a variety of mechanisms. A hot carrier injection is a possibility since a high voltage is applied to the drain of the drift transistor after gate-oxide breakdown. The triggering of the snapback N-P-N transistor may also explain the high bulk current. These hypotheses are tackled by performing bulk current measurements under a high voltage stress.
The drift transistor corresponding to the bitcell tested in figure 5.2 is characterized separately using a DC voltage ramp applied to the drain. The bitline and bulk current measurements are plotted in figure 5.5.

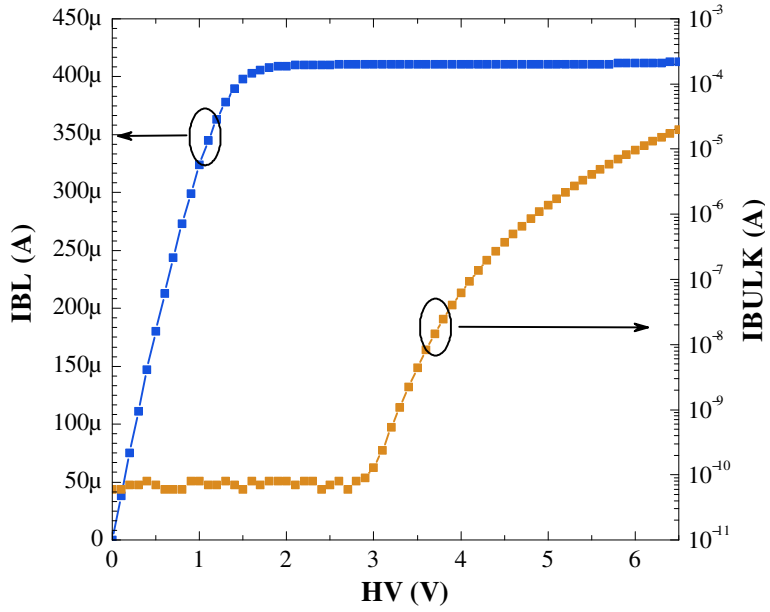


Figure 5.5: BL and bulk current measurements under a DC voltage ramp performed on a single drift transistor fabricated in a logic 45-nm CMOS process.

The characteristic of the bitline current is, as expected, following the behavior of a transistor in a saturation mode. The bulk current starts to increase for a drain voltage higher than 3V. The maximum value is about 20μA. From this simple

characterization, a hot carrier injection mechanism may not in itself explain a bulk current overshoot. Indeed, the overshoot amplitude shown in figure 5.2 reached 700μA whereas the maximum bulk current from the single drift transistor is 20μA for the same device dimensions.
In addition to DC characterizations, high voltage pulses were applied to the drain of the drift transistor. The bitline and the bulk current were measured using an oscilloscope as shown in figure 5.1(a). No particular overshoot was observed on a single drift transistor.
This hypothesis consisting in turning on the snapback N-P-N transistor is also discarded because the sign of the bulk current is not compliant with the operating condition of such a device. The bulk current goes out of the antifuse bitcell whereas the base current of a N-P-N transistor should flow from the base to the emitter, i.e. from the bulk to the source of the drift transistor.

5.1.1.2 Single capacitor

The measurements performed on a single drift transistor were not successful to explain the bulk current overshoot occurring during the programming operation of a drift antifuse bitcell. Characterizations were performed on single capacitors. In fact, only matrices of 50 capacitors connected in parallel were available for testing. A schematic and a cross-section of the test structure are depicted in figure 5.6.

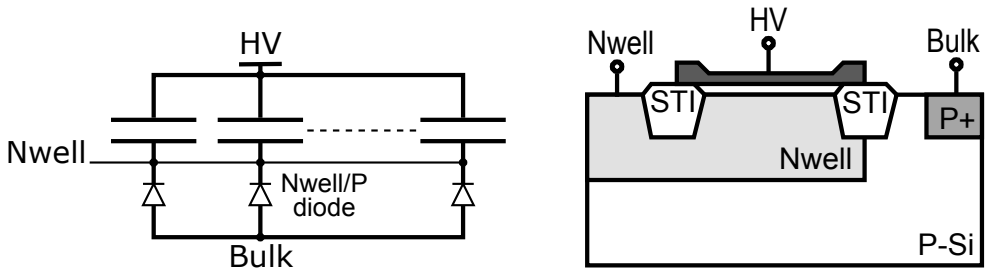


Figure 5.6: Schematic and cross section of an antifuse capacitor including body diodes. The schematic illustrates a structure comprising 50 antifuse capacitors connected in parallel.

Due to the reverse-biased Nwell/Psubstrate junction, no reasonable path allowing a high current can be identified. Except a leakage current from this junction, most of the current should flow from HV to Nwell through the broken dielectric.
The matrix of antifuse capacitors was characterized under a constant voltage stress of 5V. The Nwell and bulk current were measured using the 50-Ω inputs of an oscilloscope as in the experimental setup depicted in figure 5.1(a).

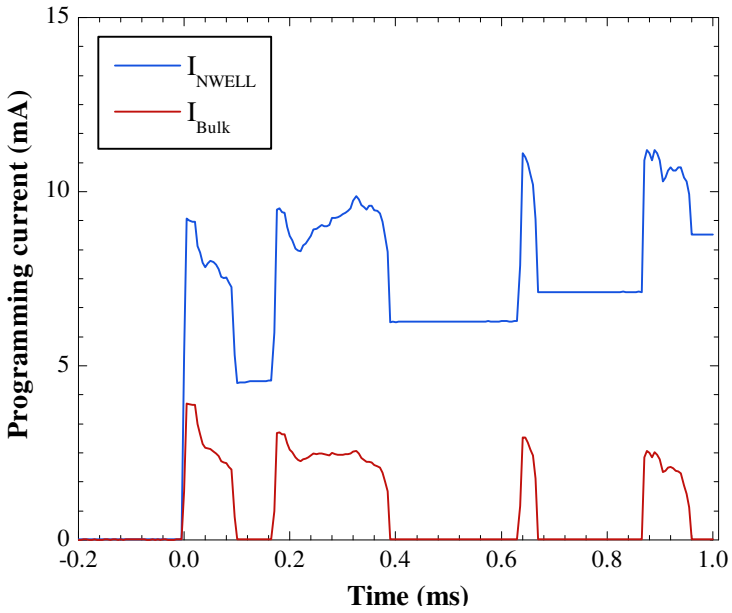


Figure 5.7: Measurements of Nwell and bulk currents on a matrix of 50 antifuse capacitors connected in parallel. The test structure was designed and fabricated in a logic 45-nm CMOS process. The stress voltage is HV=5V.

Like for an antifuse bitcell, the gate-oxide breakdown in an elementary single antifuse capacitor is clearly reflected by a steep increase in current. The Nwell current reaches almost 10mA because no device limits the programming current. Then, the current decreases down to 5mA with a particular behavior. Due to the matrix test structure, successive breakdown events occurred because the capacitors failed one after another. This is the reason why the average Nwell current increases after a capacitor breakdown.

The gate-oxide breakdown is also seen on the bulk current waveform. A maximum amplitude of 4mA can be noticed during the transient events following the dielectric breakdown. Furthermore, the trend on the transient evolution of the bulk current is similar to the Nwell current. However, the amplitude is lower.

Considering the device structure depicted in figure 5.6, a current of 4mA crosses a reverse-biased NWell/P-substrate junction. The root cause of the bulk current overshoot is likely connected to this mechanism.

5.1.1.3 Conclusion

The particularity of the drift antifuse bitcell architecture is the connection of a transistor and a capacitor using a Nwell. As a consequence, a bulk current can be generated either from the channel of the access transistor or from reverse-biased junctions.

Characterizations were performed on single drift transistors and antifuse capacitors. The hot carrier bulk current from the access transistor did not match the overshoot amplitude measured on antifuse bitcells. This hypothesis was therefore discarded. Moreover, the triggering of the snapback bipolar transistor could not be achieved by applying the programming voltage condition on the drift transistor, neither in a DC nor a transient mode.
The antifuse capacitor is composed of a N-type polysilicon gate, a silicon dioxide dielectric and a Nwell in a P-subtrate silicon. According to the voltage programming conditions, the bulk current must cross a reverse-biased junction to flow in the bulk electrode.
Programming operations were performed on single antifuse capacitors. A high bulk current is observed that a reverse-biased P-N junction cannot explain. Contrary to the bitcell, the programming current is not limited by the access device. A maximum amplitude up to 4mA was reached whereas 700μA was obtained in a bitcell. This difference can be explained by the uncontrolled post-breakdown current. It emphasizes a gain between the Nwell current and the bulk current. Since the Nwell current is limited by the saturation of the drift transistor, the bulk current is lower in a drift antifuse bitcell than in a sole antifuse capacitor.
As a first result, the cause of this side effect is connected to the antifuse capacitor architecture. Further characterizations are performed to study the influence of programming parameters such as voltage, current and temperature.

5.2 Characterizations of the phenomenon

The role of the antifuse capacitor on the bulk current overshoot was emphasized in preliminary studies and experiments presented in section 5.1. At first sight, the mechanism appeared as triggered by the gate-oxide breakdown as the overshoot is synchronized with the steep bitline current increase. Then, the bulk current decays for at least hundreds of microseconds. The mechanism was therefore noticed only during the post-breakdown phase.
This section aims at presenting different characterization methodologies and results.

5.2.1 DC characterizations

The contribution of the antifuse capacitor and the drift transistor are further studied by performing DC characterizations on an antifuse bitcell, a single capacitor and a single drift transistor. Bitline and bulk currents of the different devices are measured under a DC voltage ramp stress up to 7V and compared. Thus, the contribution of the capacitor and the drift transistor can be evaluated.

Devices were designed and fabricated in a logic 45-nm CMOS process. The dimensions of the single capacitor and the drift transistor are the same as the 16-μm^2 antifuse bitcell listed in table 5.1.
Bitline currents for the three devices are plotted in figure 5.8.

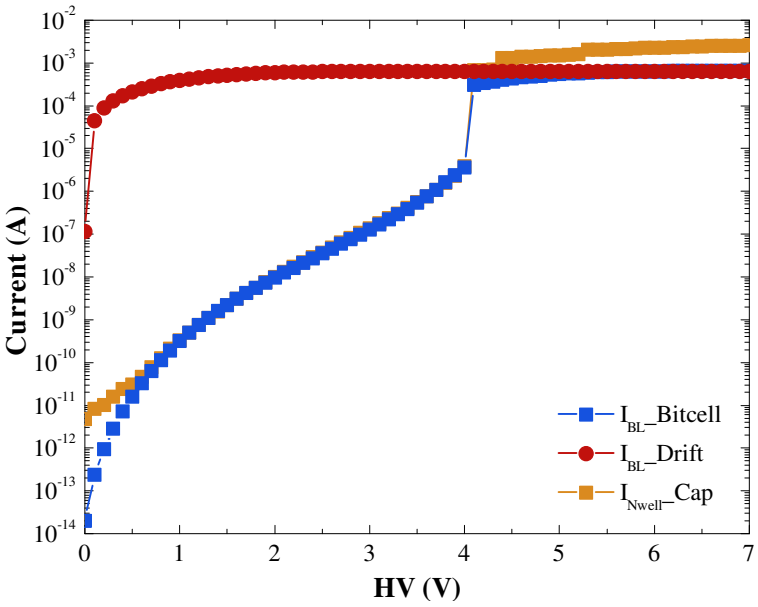


Figure 5.8: Bitline/Nwell current measurements performed on a single antifuse bitcell, drift transistor and antifuse capacitor under a DC voltage ramp.

As presented in section 3.4.2, the DC characterization allows the study of the pre-breakdown and post-breakdown phases. However, the transient aspects are out of the scope of this methodology.
The bitline current of the antifuse bitcell is increasing during the wearout phase due to tunneling mechanisms occurring in the gate-oxide of the capacitor as detailed in chapter 4. The breakdown event occurs at HV=4V. Then, the bitline current reaches the saturation level of the drift transistor.
The Nwell current of the single antifuse capacitor is equal to the bitcell current in the wearout phase until the breakdown event. The post-breakdown current is not limited by any access device and reaches a maximum value of roughly 2mA. This compliance is likely due to the stray resistance of the electrodes and the back-end of the test structure.
The contributions of the antifuse capacitor and the drift transistor are clearly emphasized. The current is driven by tunneling mechanisms in the dielectric before breakdown whereas the bitline current is limited by the saturation of the drift transistor in post-breakdown.
The bulk currents in the three devices were also measured and compared as in figure

5.9.

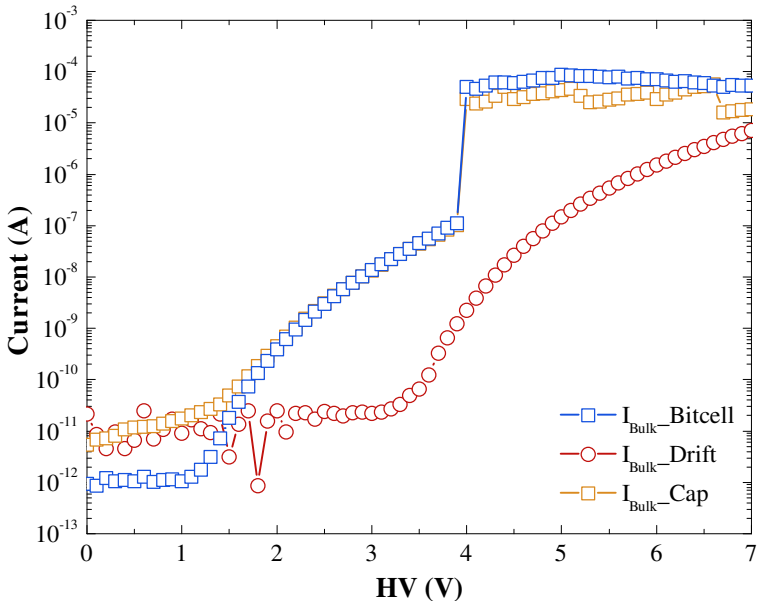


Figure 5.9: Bulk current measurements performed on a single antifuse bitcell, drift transistor and antifuse capacitor under a DC voltage ramp.

As mentioned in 5.1, the bulk current of the drift transistor caused by hot carrier injection is much lower than the overshoot amplitude. Hence, this hypothesis was discarded. The comparison between the bitcell and drift transistor DC current leads to the same conclusion as I_{bulk}_Drift does not match I_{bulk}_Bitcell.

The bulk current in the capacitor and the bitcell follows the same trend for HV from 1.5 to 4V. A maximum amplitude of 100nA is measured. Even though the bulk current seems negligible, this observation must be counterbalanced with the low voltage amplitude compared to the programming condition.

An increase of almost three orders of magnitude can be noticed at breakdown on the bitcell and single capacitor bulk current. The amplitude is steady from 4V to 7V and is slightly higher for the bitcell. There is a significant difference between the amplitude measured during transient and DC characterizations (see in figures 5.2 and 5.7). Thus, the observation and the study of this side effect phenomenon necessitates different experimental approaches in order to explore the phase between the transient and DC experimental conditions.

The significant DC bulk current measured along the wearout phase may be connected to the overshoot mechanism. Hence, the study of the phenomenon is not solely restricted to the post-breakdown phase.

5.2.2 Impact of programming conditions

The impact of the programming and wordline voltage on the bulk current overshoot is investigated. Programming operations were performed under different programming voltage amplitude. In addition, the wordline voltage was changed in order to modify the saturation current of the drift transistor, i.e., the bitline current.

5.2.2.1 Programming voltage

Programming operations were performed under various programming voltage amplitudes. The bitline and the bulk current measurements were performed on 16-μm^2 antifuse bitcells fabricated in a logic 45-nm CMOS process as listed in table 5.1. Example of characterizations are shown in figure 5.10.

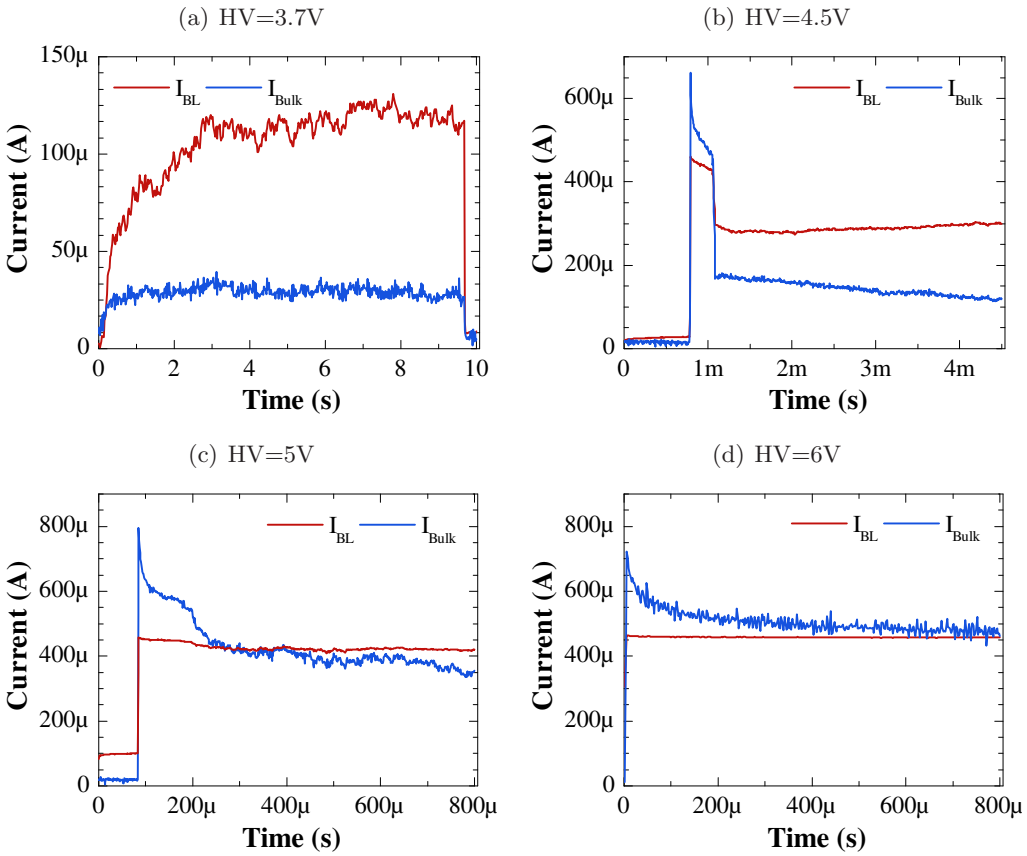


Figure 5.10: Measurements of bitline and bulk currents performed on drift antifuse bitcells designed and fabricated in a logic 45-nm CMOS process for various HV amplitudes.

The measurements performed under a programming voltage of 3.7V are shown in figure 5.10(a). Due to the long time-to-breakdown, the experimental time window is set at 10s.

As presented in chapter 4, a progressive breakdown occurs under such a low voltage. The maximum current reached is much lower than the saturation voltage of the drift transistor. This observation emphasizes that the programming voltage is not sufficient to saturate the access device.
No bulk current overshoot is measured. Like the bitline current, the increase is progressive and the maximum amplitude is much lower than the value measured under high voltage amplitudes.
The breakdown mode is different in figure 5.10(b). A hard breakdown arises after 800μs reflected by a steep increase on the bitline and the bulk currents. Contrary to previous observations, the bitline current is not steady during the post-breakdown phase. In fact, it reaches the saturation level of the access device right after the breakdown event. At the same time, the bulk current rises up to 650μA. Then, both currents decreases slowly for 300μs until a steep decay down to 300μA for the bitline current and down to 200μA for the bulk current.
This change in bitline current reflects an evolution of the drain voltage of the drift transistor.
For a higher programming voltage (HV=5V), the bulk current shape shown in figure 5.10(c) is similar to the waveform obtained for HV=4.5V. After the gate-oxide breakdown, I_{bulk} decays slowly until a steeper decrease. The bitline current is not perfectly steady. Hence, the drift transistor is at the edge of saturation.
The post-breakdown phase was previously discussed for HV=6V (see in figure 5.2), The bitline current is steady and the bulk current decreases slowly down to 450μA.
The measurements shown in figure 5.10 emphasize an impact of the programming voltage amplitude on the bulk current overshoot shape. However, the bitline current is also impacted when the drift transistor is not saturated.
Furthermore, the bulk current amplitude measured after 800μs under HV=6V does not match the amplitude measured in a DC mode. Indeed, a steady amplitude of 50μA was measured in the post-breakdown phase. This observation leads to the hypothesis of a transient phase between the breakdown event and the DC bulk current. The time scale of this period is longer than hunderds of microsecond. As a consequence, the influence of the programming voltage on the bulk current overshoot mechanism is also studied during a longer time scale. The bitline and the bulk currents are measured in the post-breakdown phase during 100ms for various programming voltage amplitudes.
Characterizations were performed on 10-μm^2 drift antifuse bitcells designed and fabricated in a logic 40-nm CMOS process as listed in table 5.1. Corresponding waveforms are plotted in figure 5.11.

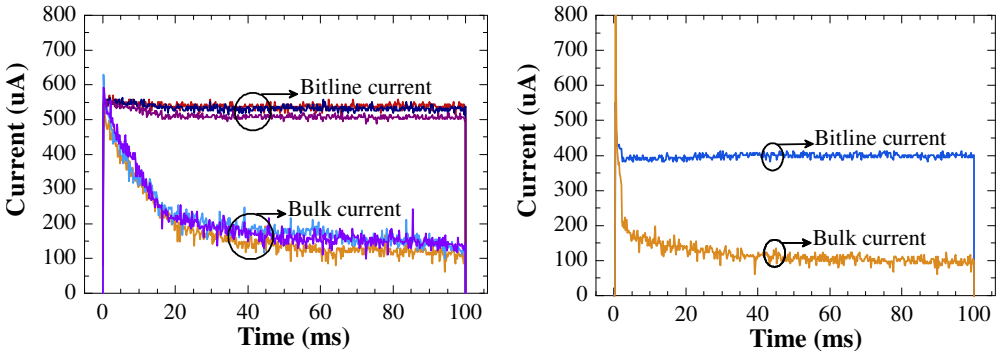


Figure 5.11: Bitline and bulk current measurements performed on drift antifuse bitcells designed and fabricated in a logic 40-nm CMOS process for various HV amplitudes during 100ms for HV=5, 5.5 and 7V (a) and for HV=4.5V (b).

The comparison of the bulk current waveforms measured under different programming voltage amplitudes emphasizes two modes.

- **The drift transistor is saturated.** For HV from 5 to 7V, the drift transistor is saturated. Hence the bitline current is the same in this programming voltage range as shown in figure 5.11(a). The bulk current waveforms are similar as well. A maximum amplitude of 700μA is reached after the gate-oxide breakdown followed by a linear decrease down to 200μA. Then, the bulk current decay is slower from 20ms to 100ms. The minimum value is roughly 150μA after 100ms.

- **The drift transistor is not saturated.** For HV $<$ 5V, the voltage applied to the drain of the drift transistor is not sufficiently high to maintain the saturation of the access device. The bitline current is therefore lower. An example of bitline and bulk current waveforms is shown in figure 5.11(b) for HV=4.5V. The bulk current decay is much faster as the value 200μA is reached in approximately 3ms. After 100ms, the amplitude is about 100μA.

The direct impact of the programming voltage amplitude on the bulk current overshoot shape is not obvious. In fact, the contribution of the programming voltage and the programming current cannot be disassociated when the drift transistor is not saturated. However, the bulk current shape did not differ for HV from 5V to 7V, a range in which the drift transistor is saturated. Antifuse bitcell are programmed in industrial products in the latter HV range.

5.2.2.2 Cumulative programming

The stability of the phenomenon was studied using a cumulative programming methodology. Pulses of 100μs width were applied to the wordline of the access transistor while the bitline current and the bulk current were measured.

The bulk current measurements are plotted in figure 5.12 and superimposed on the same graph. Each waveform corresponds to a programming pulse applied to the wordline. Measurements were performed on a 10-μm^2 antifuse bitcell designed and fabricated in a logic 40-nm CMOS process as listed in table 5.1.

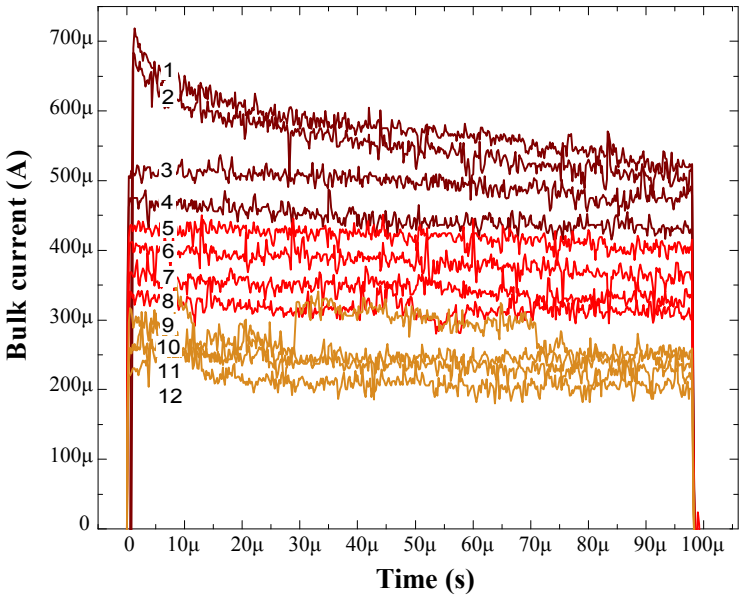


Figure 5.12: Bulk current measurements during 12 successive programming pulses. HV=6V.

The antifuse bitcell is programmed during the first programming pulse. Due to the high programming voltage (HV=6V), the wearout phase duration is about 1μs. As expected, the bulk current overshoot reaches 700μA and decays down to 550μA. The evolution of the bulk current is similar during the second programming pulse. The reduction of the bulk current amplitude is clearly seen starting from the third programming pulse. Then, a linear decay of tens of micro amps is noticed during the next programming pulses. A bulk current of 200μA was reached after twelve pulses.

Cumulative programming emphasized that the mechanism in not solely triggered by the breakdown event. A significant bulk current is measured even after twelve programming pulses. However, the mechanism appears attenuated.

5.2.2.3 Programming current

The impact of the saturation current of the drift transistor is studied in the following experiments. Antifuse bitcells were programmed under HV=6V while the worldline driving signal amplitude was changed. Thus, one bitcell was programmed with WL=2V and another with WL=1.5V. The resulting saturation current amplitudes 450μA and 240μA respectively.

Characterizations were performed on 16-μm^2 drift antifuse bitcells designed and fabricated in a logic 45-nm CMOS process as listed in table 5.1.

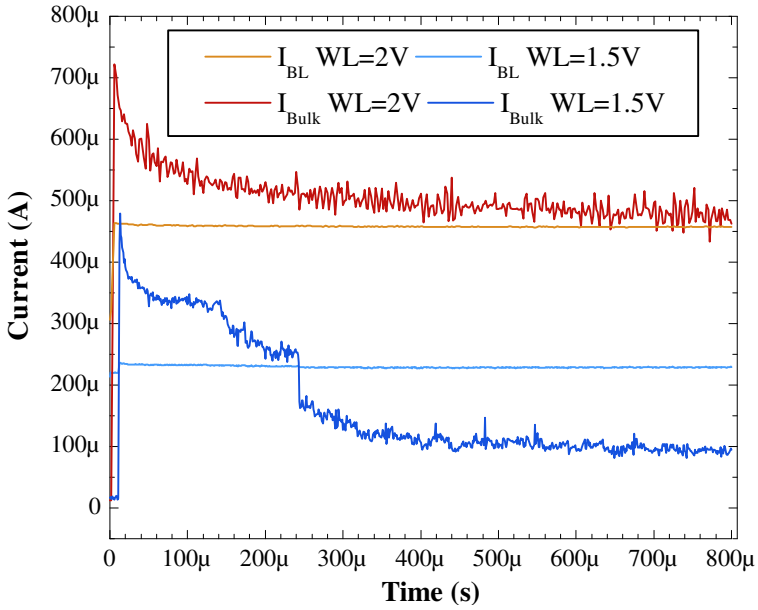


Figure 5.13: Bitline and bulk current measurements performed on drift antifuse bitcells designed and fabricated in a logic 45-nm CMOS process for various V_{WL} amplitudes.

The shape of the bulk current obtained for HV=6V and WL=2V was previously presented in figure 5.2 and 5.10(d). The bitline and bulk currents are compared to the waveforms measured for WL=1.5V.

As mentioned previously, the maximum amplitude of the bulk current overshoot is slightly lower than 500μA whereas 700μA was reached for WL=2V. The shape is also different. The mechanism appears attenuated by a lower saturation current of the drift transistor. A current 100μA is obtained after 800μs for WL=1.5V whereas I_{bulk} is slightly lower than 500μA for WL=2V after the same duration.

As presented in the latter characterization, the bitline current amplitude has a first-order impact on the bulk current overshoot as the higher the saturation current of the drift transistor, the higher the overshoot amplitude. Furthermore, the phenomenon seems attenuated by a lower bitline current.

5.2.2.4 Summary & conclusion

The side effect leading to a bulk current overshoot during a programming operation has been characterized under various conditions. The phenomenon was measured for 100ms with variations from 800μA to 100μA. Moreover, the maximum amplitude is obtained right after the gate-oxide breakdown. The power consumption is therefore much higher for the first 100μs of the post-breakdown phase. The bulk current overshoot has therefore a severe impact on the antifuse memory.

The influence of two main parameters was studied in this section. First, programming operations were performed on antifuse bicells under different programming voltage amplitudes. It was observed that the bulk current was not impacted if the drift transistor remained saturated in the post-breakdown phase. Consequently, the bulk current shape is the same for a HV range from 5 to 7V. For HV $<$ 5V, the drift transistor is not saturated after the breakdown event. The bitline current is not constant and the bulk current is attenuated. The dependence in programming voltage amplitude seems connected to the operation regime of the drift transistor. Second, experiments were performed using different worldline signal amplitude in order to change the saturation current of the drift transistor. It turned out that the bulk current was strongly impacted by this condition as the lower I_{dsat}, the lower I_{bulk}.

The dependence in temperature was studied. There was no particular impact apart from the saturation current of the drift transistor. The phenomenon was neither amplified nor attenuated.

The impact of the bulk voltage was also experimented. An influence of the body effect was noticed. However, it was not possible to disable the mechanism.

Parameters	**Impact**	**Trend**
Wordline voltage	Overshoot amplitude and duration	Higher V_{WL}, higher I_{dsat}, higher I_{bulk}
Programming voltage	No direct influence unless the drift transistor enters in a linear regime	Lower I_{bulk} if $I_d < I_{dsat}$
Temperature	No direct influence	I_{bulk} follows I_{dsat} variations
Bulk voltage	No direct influence	I_{bulk} follows I_{dsat} variations

Table 5.2: Summary of the impact of programming parameters on the bulk current overshoot.

5.3 Analysis of assumptions on the root cause

The influence of the programming conditions on the bulk current overshoot were investigated in the previous section. However, the root cause of this side effect could not be pointed out. Since the phenomenon was noticed on various CMOS technology, e.g. 55-nm, 45-nm and 40-nm, and various bitcell dimensions (see in table 5.1), it can be concluded that this side effect is persistent and will likely occur in future CMOS technologies. Investigating a root cause appears therefore essential in order to understand and to try to disable the underlying mechanism. The goal is to point out the most probable assumption, work on a verification and provide a solution to overcome the overshoot issue.
Different analyses are detailed in this section.

5.3.1 Electron and hole transport

Since the bitline and the bulk current have a particular evolution before and after the gate-oxide breakdown, it is worth to investigate the transport of holes and electrons within a drift antifuse bitcell during the wearout and post-breakdown phases.
A cross section of a drift antifuse bitcell designed in a logic 55-nm CMOS process was generated using a Technology Computer-Aided Design tool [1]. A similar bulk current overshoot was observed in this technology as in the characterization of the 45-nm and 40-nm CMOS devices. The simulation work was performed by Elise Le-Roux [2] and François Wacquant [3].

5.3.1.1 Wearout phase

The wearout phase is illustrated in figure 5.14.

[1] Synopsis, Sentaurus Workbench advanced. Version:D-2010.03
[2] Elise Le-Roux is with STMicroelectronics/TR&D/CCDS/Fuse solutions.
[3] François Wacquant is with STMicroelectronics/TR&D/STD/Technology modeling.

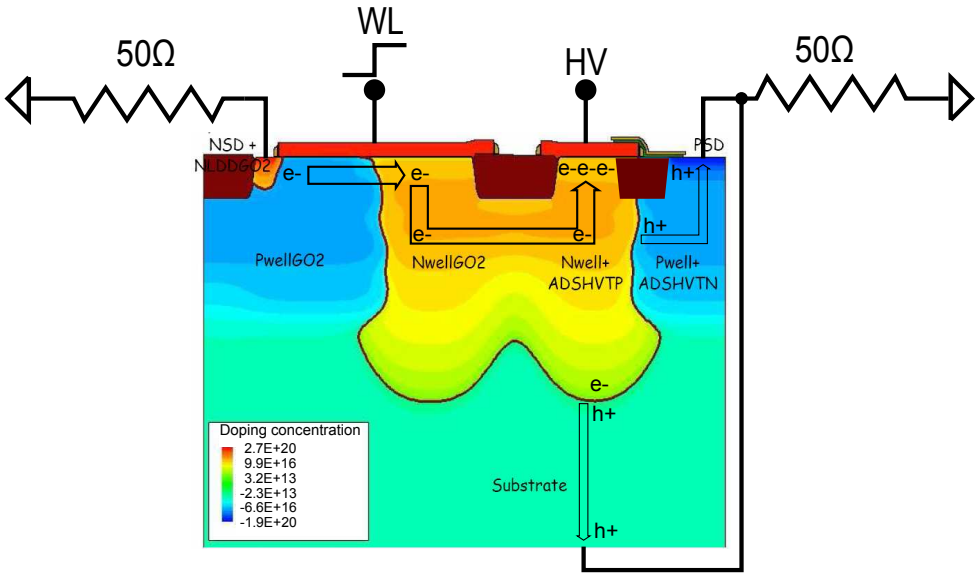


Figure 5.14: Cross-sectional view of an antifuse bitcell designed in a logic 55-nm CMOS process and the experimental conditions during the wearout phase.

As presented in figure 5.1, the bitline and the bulk currents are measured using 50-Ω inputs of an oscilloscope. A DC high-voltage is applied to the gate of the antifuse capacitor while the voltage pulse is applied to the gate of the drift transistor. As a consequence, the electrons flow from the source to the Nwell through the channel of the access transistor and are thus accumulated beneath the gate-oxide of the antifuse capacitor. As discussed in chapter 4, the electrons have sufficient energy to cross the gate-oxide and to damage the dielectric material until the gate-oxide breakdown.

A significant bulk current was measured before the gate-oxide breakdown as shown by DC characterizations in section 5.2.1. Furthermore, it was shown experimentally that the bulk current is due to the antifuse capacitor. Considering the structure of the drift antifuse bitcell, the bulk current is, in fact, a hole current (majority carriers in the bulk P-layer). The Nwell/P-substrate junction is reverse-biased along the wearout phase. As detailed in chapter 4, the Nwell voltage is equal to $I_{wearout} \cdot R_{on}$ and is lower than 1V. This condition is not sufficient to trigger an avalanche mechanism.

TCAD simulations were performed on an antifuse capacitor designed in a logic 55-nm CMOS process in order to evaluate the energy-band bending and the resulting electron and hole flows in the device. An energy-band diagram for HV=1.2, 3.6 and 6V is plotted in figure 5.15.

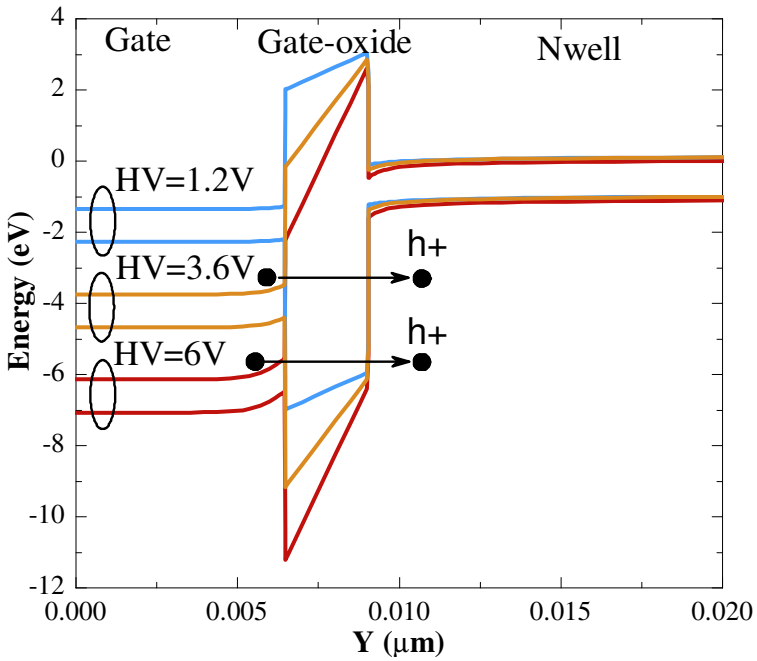


Figure 5.15: Energy-band diagram of an antifuse capacitor designed in a logic 55-nm CMOS process.

As presented in chapter 3, the barrier energy-band bending changes according to the voltage amplitude. For HV=6V, the barrier is triangular as in a Fowler-Nordheim conduction mode whereas a direct tunneling mechanism is involved in the electron conduction for HV=1.2V and 3.6V.

The electron current is usually considered in both mechanisms due to the difference in barrier height for the electrons and holes ($\phi_e = 3.1$eV, $\phi_h = 4.6$eV). The carriers transport from the conduction band from the silicon substrate to the gate is computed. However, the bulk current amplitude computed in TCAD simulations was not consistent with the silicon measurements.

The programming voltage of an antifuse capacitor is much higher than the traditional range studied in semiconductor physics. Indeed, the conduction of holes is usually neglected due to a higher barrier height. Since the barrier shape is also dramatically impacted by the programming voltage range, a hole current flowing in the bulk is not negligible. This is the reason why a band to band tunneling was assumed in the simulation model. Simulations emphasized that a direct tunneling of holes from the conduction band of the gate to the valence of the silicon substrate is possible for a voltage higher than 1.5V.

Gate and bulk currents are plotted in figure 5.16. The type of carriers, i.e. electrons and holes are identified.

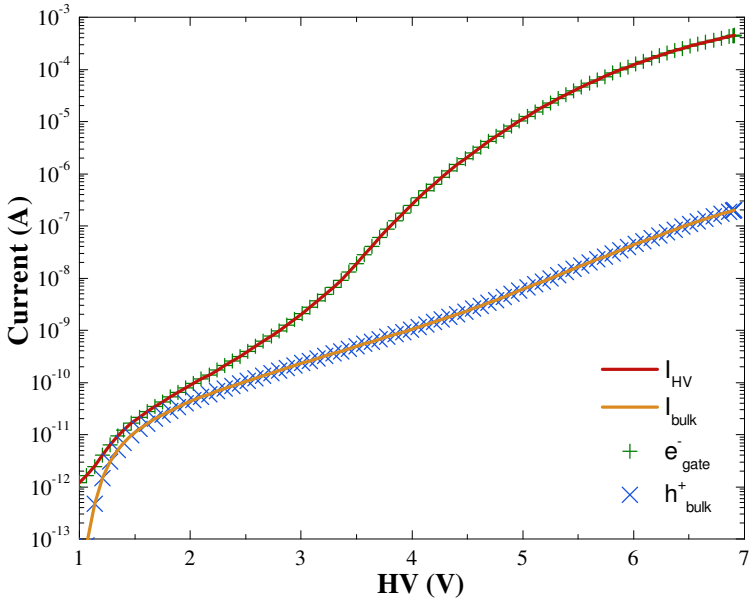


Figure 5.16: Gate and bulk current of an antifuse capacitor designed in a logic 55-nm CMOS process.

The simulated currents have a similar amplitude than the measurements performed on an antifuse capacitor designed and fabricated in a logic 55-nm CMOS process. In fact the order of magnitude between I_{HV} and I_{bulk} is correct. A finer tuning of the models would yield more accurate results. However, those qualitative results allow an assumption on the origin of the bulk current during the wearout phase. A band-to-band mechanism seems relevant. Very recent TCAD simulations confirmed that minority carriers (holes) are emitted by the N-type polysilicon gate of the antifuse capacitor and are transported to the silicon substrate by a band-to-band tunneling mechanism.

5.3.1.2 Post-breakdown phase

Most of the concerns regarding the bulk current overshoot are centralized on the post-breakdown phase during which the programming current is severely impacted. After the gate-oxide breakdown, the bulk current amplitude is similar or even higher than the bitline current. Like for the wearout phase, it is worth to study the electron and hole flow in the antifuse device.
The post-breakdown phase is illustrated in figure 5.17.

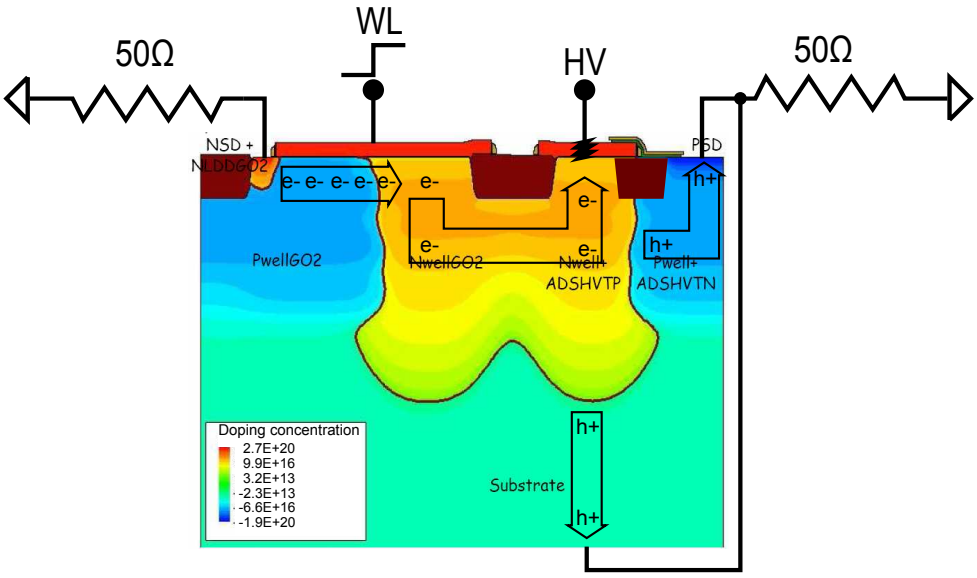


Figure 5.17: Cross-sectional view of an antifuse bitcell designed in a logic 55-nm CMOS process and the experimental conditions in a post-breakdown phase.

The bitline current is limited by the drift transistor. Hence, the electrons flow from the source of the access device to the Nwell and then by the breakdown spot.
The path of the high bulk current rises more questions. After the gate-oxide breakdown, the Nwell voltage, i.e. the drain voltage of the drift transistor, increases significantly as the drift transistor enters in a saturation regime. The Nwell/P-substrate junction is therefore reverse-biased.
A high bulk current can be generated from a reverse-biased P-N junction only if an avalanche mechanism occurs. It is commonplace to assume that this phenomenon is triggerred by a high voltage. Spice simulations were performed on a reversed-biased Nwell/P-substrate diode. The device was chosen in a 40-nm CMOS process and dimensioned according to the 10-μm^2 antifuse bitcell characteristics listed in table 5.1. A DC voltage ramp was applied to the cathode, i.e. the Nwell while the current was probed. The simulation result is plotted in figure 5.18.

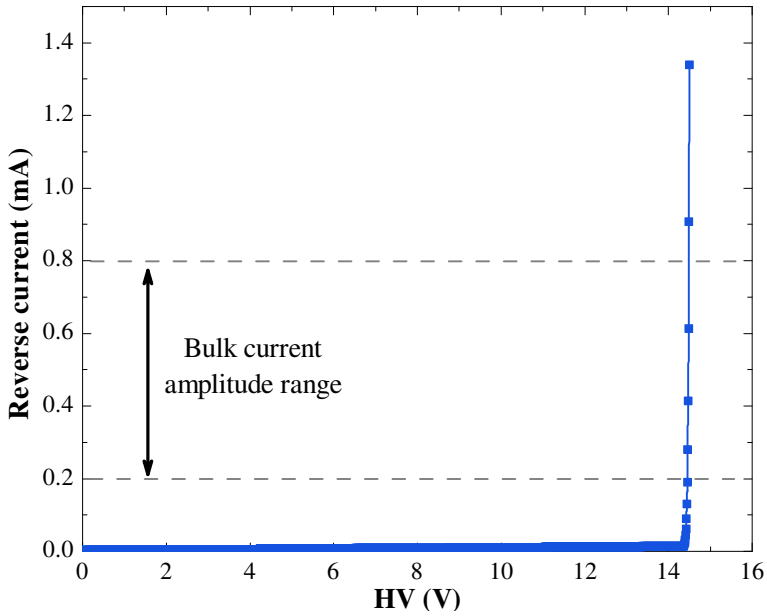


Figure 5.18: Spice simulation of a reverse-biased Nwell/P-substrate diode in a 40-nm CMOS process. The dimensions are set as in a 10-μm^2 40-nm drift antifuse bitcell (see in table 5.1).

The avalanche is clearly seen for a voltage amplitude slightly higher than 14V. Since the maximum programming voltage of a 40-nm antifuse bitcell is 7V, the avalanche mechanism cannot be triggered.
The transport of electrons and holes in an antifuse capacitor was investigated in the wearout and post-breakdown phase. TCAD simulations were performed and shown that the gate-current is mostly carried by electrons while the bulk current is carried by holes. As measured on antifuse devices, the bulk current is much lower than the HV or BL current in the wearout phase.
The modeling of the breakdown spot is not in the scope of conventional TCAD simulators. However, it can be assumed that the high bulk current is also carried by holes in the post-breakdown phase. The Nwell/P-substrate is reverse-biased. The assumption of avalanche was discarded as a voltage higher than 14V is needed to obtain a current of hundreds of micro-amp from a reverse-biased junction.

5.3.2 Parasitic bipolar transistor

According to the operation principle of a bipolar transistor, it is possible to allow a high current through a reverse-biased P-N junction. Considering the Nwell/P-substrate junction as the base-collector, a P-N-P bipolar transistor can be identified in the antifuse capacitor. A cross-sectional view of an antifuse bitcell is depicted in figure 5.19.

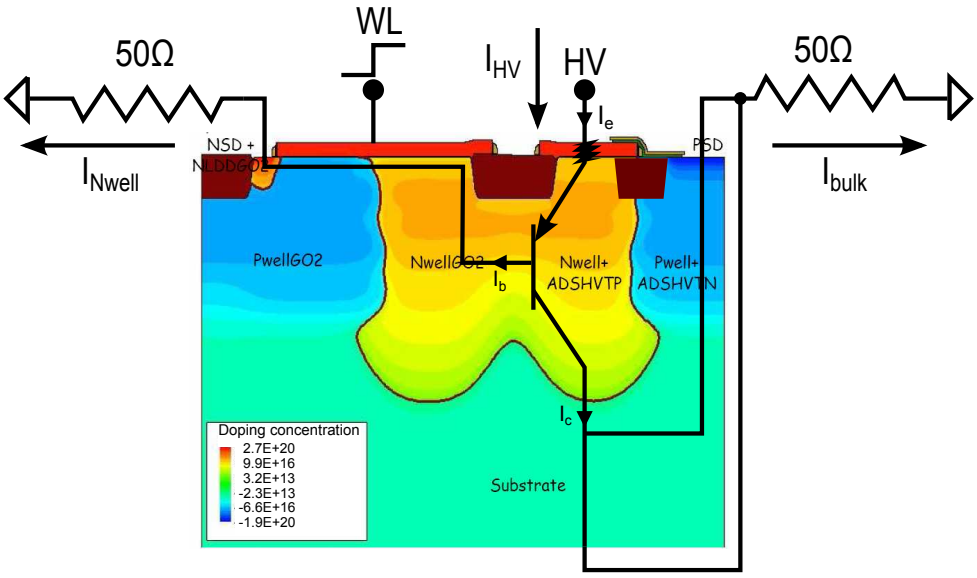


Figure 5.19: Cross-sectional view of an antifuse bitcell designed in a logic 55-nm CMOS process and the identification of a parasitic P-N-P transistor.

The base of the identified P-N-P bipolar transistor is the Nwell. The base/collector junction is therefore formed by the P-substrate. As its name suggests, the emitter/base junction should be also formed by P-type semiconductor. The structure of the antifuse capacitor is therefore not similar to a P-N-P transistor. In this study, the emitter is assumed to be the polysilicon gate and the breakdown spot.
The pertinence of this hypothesis is discussed in the following sections. The sign of the currents and the polarities in the antifuse bitcell are compared with the required operating conditions of a P-N-P transistor. Then, the electron and hole flows in a P-N-P transistor are detailed and compared with the programming operation of a drift antifuse bitcell.

5.3.2.1 Current signs and polarities

A bipolar junction transistor (BJT) is a three-terminal device used for amplification or switching applications. Unlike a field-effect transistor which is an unipolar device, the operation of a BJT involves both electrons and holes as carriers. The structure is composed both of N-type and P-type semiconductors. There are two types of bipolar transistors such as N-P-N and P-N-P. The structure and the symbol of a P-N-P transistor are depicted in figure 5.20.

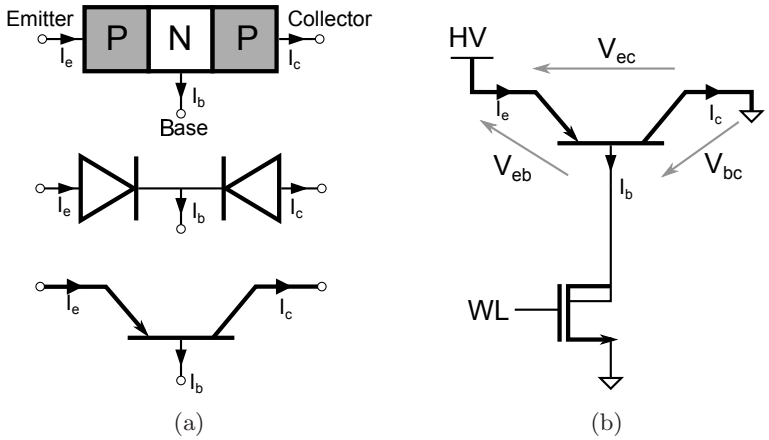


Figure 5.20: Structure and symbols of a P-N-P transistor (a). P-N-P transistor connected to a drift access transistor (b).

The structure of a P-N-P transistor can be approached by a two-diode model composing a base/emitter and a base-collector P-N junction. The device is turned on by forward biasing the base/emitter junction and by applying a voltage sufficiently high between the emitter and the collector. As a consequence, there is a base current flowing through the forward-biased emitter/base junction while a collector current flows through the reverse-biased base/collector junction. Since a bipolar transistor is a three-terminal device, the emitter current is:

$$I_e = I_b + I_c \quad \text{or} \quad I_{HV} = I_{Nwell} + I_{Bulk} \tag{5.1}$$

As mentioned above, the conduction mode of a P-N-P transistor depends on the biasing of the emitter/base and collector/base junctions. The required voltage conditions for the off-state, saturation and linear regime are given in table 5.3.

	Off	**Saturation**	**Linear**
V_{eb}	$< V_t$	$> V_t$	$> V_t$
V_{ec}	-	$< V_{ecsat}$	$> V_{ecsat}$

Table 5.3: Modes of operation of a P-N-P bipolar transistor and voltage conditions

The signs of the terminal currents are in accordance with the operation of a P-N-P transistor. As sketched in figure 5.20(b), the emitter current flows out of the high voltage source. The anode of the emitter/base junction is connected to HV while the cathode is connected to the drain of the drift MOS transistor. Thus, the junction is forward biased.
The sign of the base current is also correct as it is controlled by the bitline current, i.e. the saturation current of the drift transistor.

Finally the collector current is assumed to be the bulk current overshoot. The measurements have confirmed the sign of the latter current. Moreover, this node is tied down to the ground. The base/collector junction is therefore reverse biased.

5.3.2.2 P-N-P structure

As a first approach, the emitter is assumed as the N-type polysilicon gate and the breakdown spot that are replaced by a P-type layer for the purpose of analysis. Since electrons and holes are involved in the operation of a P-N-P transistor, it is worth to study the flow of carriers within the antifuse capacitor as in figure 5.21.

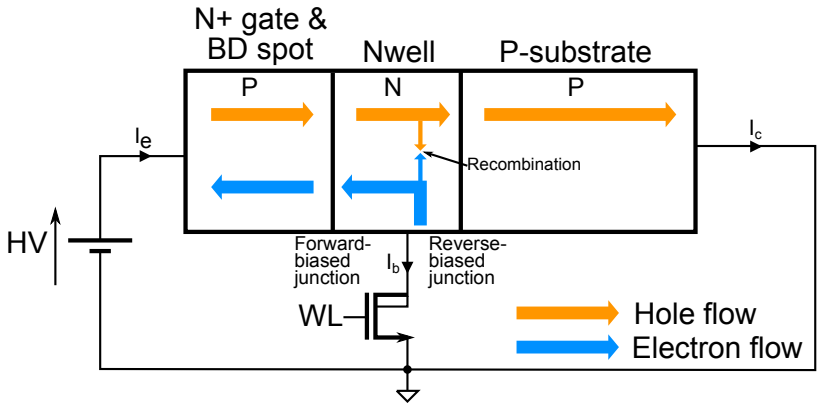


Figure 5.21: P-N-P transistor connected to a drift transistor. The gate-oxide is assumed broken.

The emitter is tied up to HV while the polarity of the base is set by the saturation voltage of the drift transistor. The emitter/base junction is therefore forward biased and the base current is, in fact, the bitline current. Electrons flow from the drift transistor to the HV source.

The forward biasing of the emitter/base junction allows the diffusion of holes in the base from the emitter interface to a depleted region near the collector. Consequently, the holes which are minority carriers in the Nwell can be either recombined with electrons or diffused through the base. Holes are accelerated by the electric field in the depleted region of the reverse-biased junction, thereby enabling the collector current.

The performance of a P-N-P bipolar transistor depends, among other criteria, on the quantity of holes recombined in the base. This condition is also linked to the existence of a hole reservoir in the N-type polysilicon gate of the antifuse capacitor. To minimize the phenomenon of recombination, the doping concentration of the base is low and the thickness must be less than the diffusion length of holes. In other words, holes are diffused through the base in less time than the lifetime of

minority carriers in the semiconductor. Due to the dimensions of the Nwell, poor performance can be expected from the parasitic P-N-P bipolar transistor.
Bipolar transistors are available in CMOS design kit although these devices are not optimized in a standard CMOS process. The structure of a conventional P-N-P transistor available in the design kit of a logic 40-nm CMOS process and an antifuse capacitor are compared in figure 5.22.

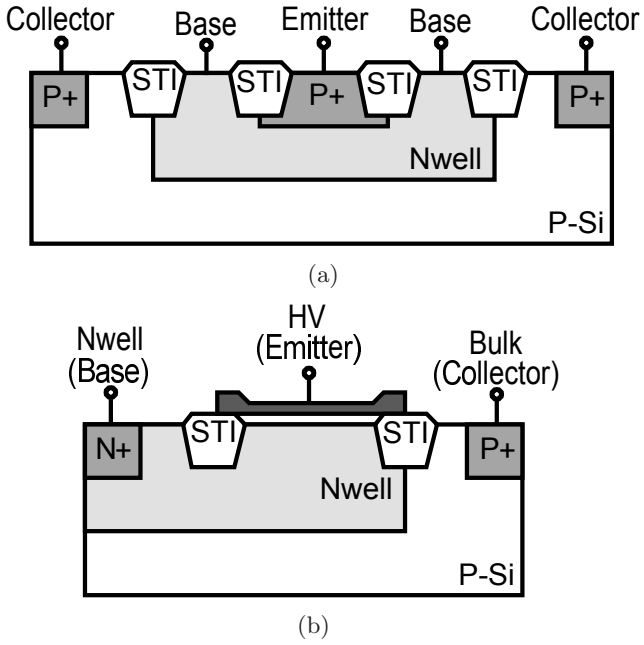


Figure 5.22: Cross-sectional view of a P-N-P transistor (a) and an antifuse capacitor (b) in a logic CMOS 40-nm process.

The base of the P-N-P transistor depicted in figure 5.22(a) is a Nwell and the collector is the P-substrate. Except the emitter, the structure of the antifuse capacitor is similar to a P-N-P transistor. The dimension and the doping concentration of the Nwell are therefore appropriate to be used as the base of a P-N-P transistor.
The TCAD simulations reported in section 5.3.1.1 have shown that the polysilicon gate is able to emit holes in the wearout phase due to a band-to-band tunneling mechanism. Even though the structure differs significantly from a proper P+ emitter, the gate can be assumed as a hole reservoir, in other words, the emitter of the P-N-P transistor.

5.3.2.3 Current-gain simulations

One of the key function of a bipolar transistor is the amplification of the base current. Thus, the collector current is amplified by a factor β. The current gain is constant if the device operates in a linear regime whereas the performance in ampli-

fication are reduced in a saturation mode. The purpose of the following simulations is to evaluate the performance of a P-N-P transistor available in the design kit of a logic 45-nm CMOS technology. However, the available devices does not accurately match the dimensions of an antifuse capacitor. The Nwell area of the simulated transistor is $12.56\mu m^2$ whereas an antifuse capacitor features $2.94\mu m^2$. Despite this difference in dimensions, the simulation results will be compared with measurements in order to emphasize the behavior of a parasitic P-N-P transistor during the programming operation of an antifuse bitcell.
The testbench used to perform current-gain simulation is depicted in figure 5.23.

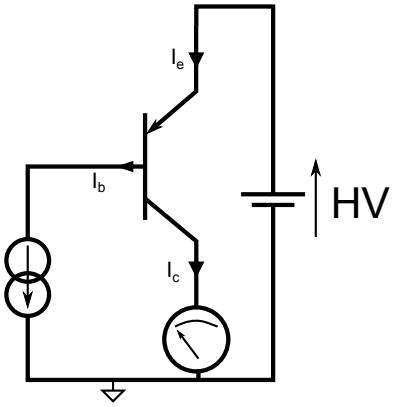


Figure 5.23: Schematic of the testbench used to simulate the gain of the P-N-P bipolar transistor. The base current is swept from 0mA to 10mA. HV=6V.

The potential of the HV source is set at 6V. A current ramp up to 10mA is applied to the base of the transistor. The collector current is measured and the gain β is calculated as:

$$\beta = \frac{I_c}{I_b} = \frac{I_{Bulk}}{I_{Nwell}} \tag{5.2}$$

Experiment 1: the experiments performed on a matrix of capacitors emphasized the high bulk current overshoot (see in section 5.1.1.2). It was also shown that the current from the Nwell was higher and exhibited a similar shape in the post-breakdown phase. These measurement are plotted in figure 5.24. The current gain was calculated as defined in equation 5.2. Measurements are plotted in figure 5.24.

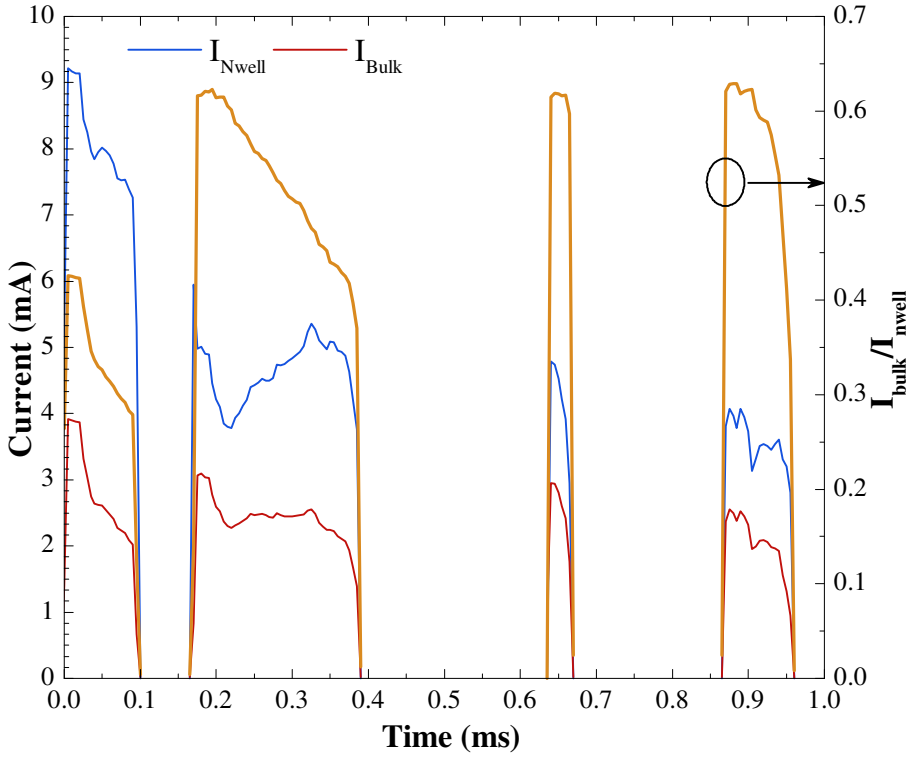


Figure 5.24: Measurement of Nwell and bulk currents on a matrix of 50 antifuse capacitors connected in parallel. The ratio I_{bulk}/I_{Nwell} is computed for the post-breakdown phase.

As mentioned previously, the Nwell current is not limited by a series drift MOS transistors. Consequently, I_{Nwell} and I_{Bulk} reaches amplitudes of several milli-amps. The calculated gain is quite low in this example. Except for the breakdown of the first capacitor, the maximum current gain is slightly higher than 0.6. It can be noticed that the gain is not constant during the post-breakdown phase.

Experiment 2: the measurements performed on antifuse bitcells for various WL amplitudes are relevant to compare the current gain (see in section 5.2.2.3). Indeed, changing the amplitude of the driving voltage of the drift transistor leads to different base current conditions. Measurements are plotted in figure 5.25. The maximum β is calculated for both conditions.

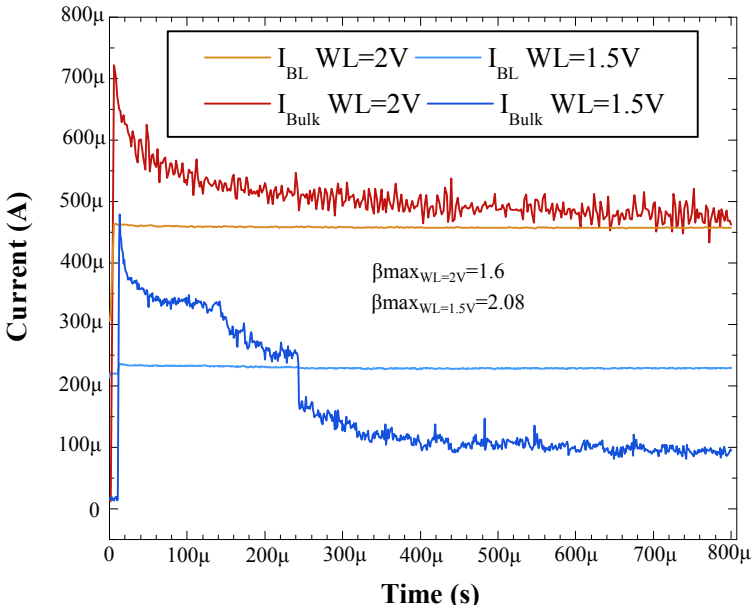


Figure 5.25: Bitline and bulk current measurements performed on drift antifuse bitcells designed and fabricated in a logic 45-nm CMOS process. The maximum β is calculated for the highest bulk current value.

The computed current-gain values are higher than the ones calculated for the capacitors. Nevertheless, a trend can be identified as the lower the base current, the higher the current gain. Furthermore, the gain is not constant during the post-breakdown phase whereas the base current is constant due to the series drift MOS transistor.

SPICE simulations: the two practical examples are compared with the simulations of the current gain performed using the test-bench depicted in figure 5.23. Simulations results are plotted in figure 5.26.

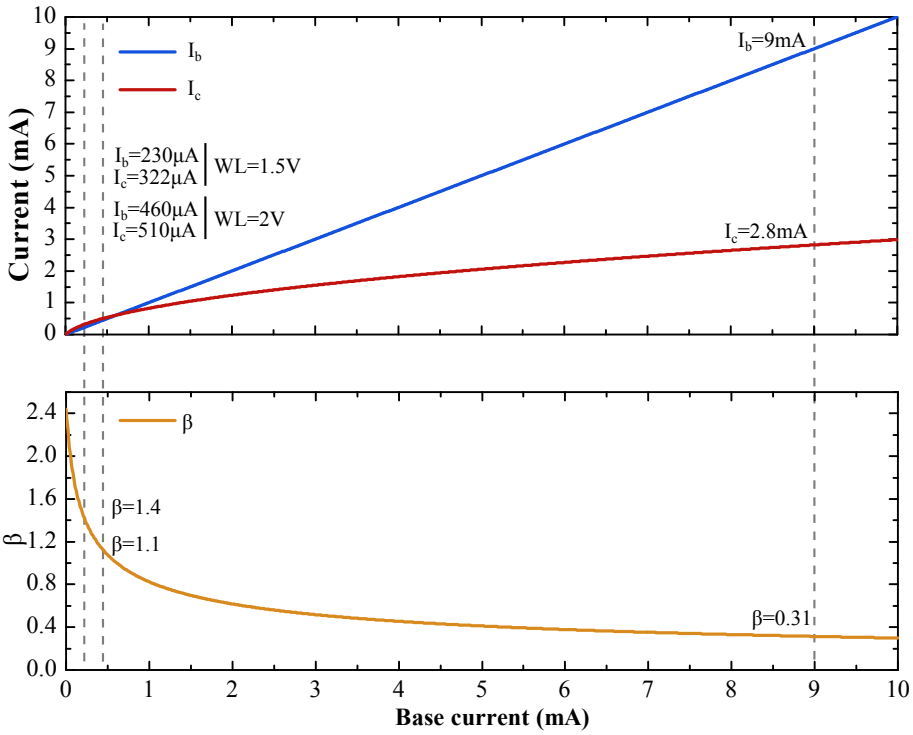


Figure 5.26: Current-gain simulations results. The collector and the gain are identified according to a base current corresponding to the conditions set in the previous experiments 1 and 2.

The gain of the transistor is not constant. An amplification of 2.4 is obtained for a low current while the gain decays down to roughly 0.3 for I_b. As expected, the performance in amplification is poor.

The trend on the evolution of the current gain is similar. A higher β is obtained for the lowest base current, e.g. 230μA. As observed for the capacitors the gain is much lower than 1 for a base current of 9mA.

The simulations of the current gain leads to the conclusions that the measurements followed the same trend. Fitting accurately the simulation and electrical characterization results is obviously challenging due to the difference in dimensions and structures between a regular P-N-P transistor and the assumed parasitic BJT device in the antifuse capacitor.

5.4 Summary and conclusion

A mechanism leading to a bulk current overshoot during the programming operation of a drift antifuse bitcell was studied in this chapter. This phenomenon was experimentally measured on various devices fabricated in logic 55-nm, 45-nm and

40-nm CMOS processes. Since this side effect has an impact on the power consumption and the performance of the antifuse memory, the understanding of the underlying physical phenomenon appears essential as well as the actions to prevent the overshoot.

A first major result is that the bulk current is clearly amplified by the failure of the dielectric.

Further electrical characterizations were performed in order to analyze the influence of various parameters such as the programming and wordline voltage amplitudes. The impact of the temperature and the bulk biasing were also investigated. A major influence of the wordline voltage was pointed out from these experiments. More precisely, a change in the driving voltage of the drift transistor leads to a different saturation current and therefore a different post-breakdown current. There is in fact a relationship between the current driven by the access transistor and the bulk current overshoot amplitude and duration. This experimental observation was confirmed by measurements performed under a variety of programming voltage amplitude. In some cases, the HV amplitude was not high enough to saturate the drift transistor in the post-breakdown phase. The impact of the temperature and the bulk biasing were also studied. Like the wordline voltage, both parameters have a repercussion on the saturation of the drift transistor, hence the bulk current amplitude.

The hypothesis of a major contribution of electrons in the bitline current and holes in the bulk current were confirmed by TCAD simulations. The study of the energy-band diagram under a high voltage stress emphasized a strong band bending due to the high programming voltage. Such extreme conditions involve a significant hole current in the wearout phase due to a band-to-band conduction mechanism. The analysis of the post-breakdown phase is more difficult because the breakdown spot cannot be modeled. Like in the wearout phase, the bitline current is carried by electrons whereas the bulk current is carried by holes. However, the Nwell/P-substrate junction which is crossed by the bulk current is reverse biased. The cause of the overshoot is probably connected to the junction. The hypothesis of an avalanche mechanism was discarded by SPICE simulations. The hypothesis of the triggering of a parasitic P-N-P transistor was then discussed. Considering the P-substrate as the collector and the Nwell as the base, a PNP transistor can be identified in the antifuse capacitor assuming that the emitter is the polysilicon gate and the breakdown spot. Since the nature of the emitter is unusual, this hypothesis was verified by studying the current signs, the polarities and the carrier transport in the bitcell. Therefore, the carrier transport in the capacitor was compared to the electrons and holes flow in a P-N-P transistor. The N-type semiconductor which

forms the base of the transistor has particular requirements for a proper operation of the bipolar transistor. The collector is the P-substrate. However, the emitter is a P-type semiconductor (P+). In other words, the emitter is a hole reservoir. In the antifuse capacitor, the emitter is assumed to be the N-type polysilicon gate and the breakdown spot. TCAD simulations have shown that holes can be emitted by the gate. The structure of a P-N-P bipolar transistor is therefore identified. Nevertheless, the current shape measured on antifuse bitcell remains unexplained. A variation of the emitter-base and emitter-colecter voltages are possible assumptions. Further studies are needed to conclude on the root cause of this mechanism. This topic is still addressed by the device staff within STMicroelectronics. Test structures have been designed and will be tested in the future. The bulk current overshoot remains a major concern.

Even though the triggering of a parasitic P-N-P bipolar transistor seems a relevant root cause, solutions to prevent the bulk current overshoot are needed. Since the programming current is not limited by the access transistor, another system is required to accurately control the current flowing through the breakdown spot. This is the reason why current compliance circuit are discussed in the next chapter.

CHAPTER 6

Post-breakdown phase and read current

After the wearout phase, the final step of the programming operation of an antifuse bitcell consists in maintaining the high voltage across the device after the breakdown event. Hence, a high current flows through the breakdown spot for a certain time. After programming, antifuse bitcells are read under a low voltage by means of measuring the leakage current of the capacitor. Since the wearout phase is rather short compared to the overall programming time, it can be concluded that most of the programming energy is consumed during the post-breakdown phase. The objective of the present study is therefore to optimize the programming sequence in order to obtain an acceptable read current amplitude while reducing the energy consumed after the gate-oxide breakdown.

This chapter focuses on the understanding of the contribution of the post-breakdown phase in the resulting amplitude of the read current. In other words, studies and experiments are performed in order to describe the degradation mechanisms of the dielectric material leading to a low resistive state.

A schematic approach of the programming operation of an antifuse bitcell is depicted in figure 6.1. The post-breakdown phase is highlighted.

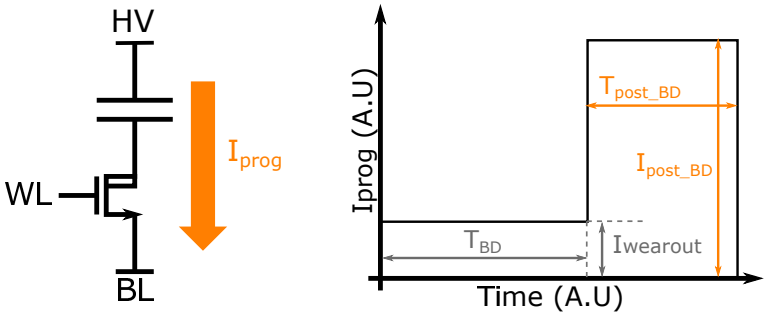


Figure 6.1: Schematic of the capacitor gate current during a programming operation.

The post-breakdown phase is defined by the post-breakdown time (T_{post_BD}) and the post-breakdown current amplitude (I_{post_BD}). The different studies reported in this chapter deal with the contribution of these two parameters on the resulting read current and on the characteristic of the breakdown path.

The read operation of an antifuse bitcell is presented in section 6.1. Experimental results introduce the electrical characteristic of a broken dielectric material, i.e. a programmed antifuse bitcell.

The effects of the post-breakdown current amplitude and the duration of the post-breakdown phase are detailed in section 6.2. For this purpose, the test structure was designed in order to accurately control the post-breakdown current amplitude.

Conclusions are drawn in section 6.3.

6.1 Read operation basics

The read operation of an antifuse bitcell is achieved by applying a voltage to the HV node and by measuring the current flowing through the device. The access transistor is turned on: WL=Vdd and BL=0V. Contrary to the programming operation, the state of the dielectric material should not be altered by the read voltage. Usually an amplitude around the operation voltage of thin-oxide MOS transistor is chosen (1.2V for a bitcell designed in a logic 40-nm CMOS process).

6.1.1 Read current distributions

The gate-oxide breakdown physics is defined as a stochastic process. Therefore, it is worth to study the statistical dispersion of the read current for a set of antifuse bitcells programmed in the same conditions. In the following example, drift antifuse bitcells designed and fabricated in a logic 40-nm CMOS process were programmed by a voltage pulse of 5.5V amplitude and 50μs duration (WL=Vdd, BL=0V). Then,

the read current was measured under a voltage of 1.2V. Dimensions of the discussed bitcells are given in chapter 4, table 4.5 ("medium" bitcell).
Read current distributions of virgin and programmed bitcells are plotted in figure 6.2. Each distribution is a set of 25 devices.

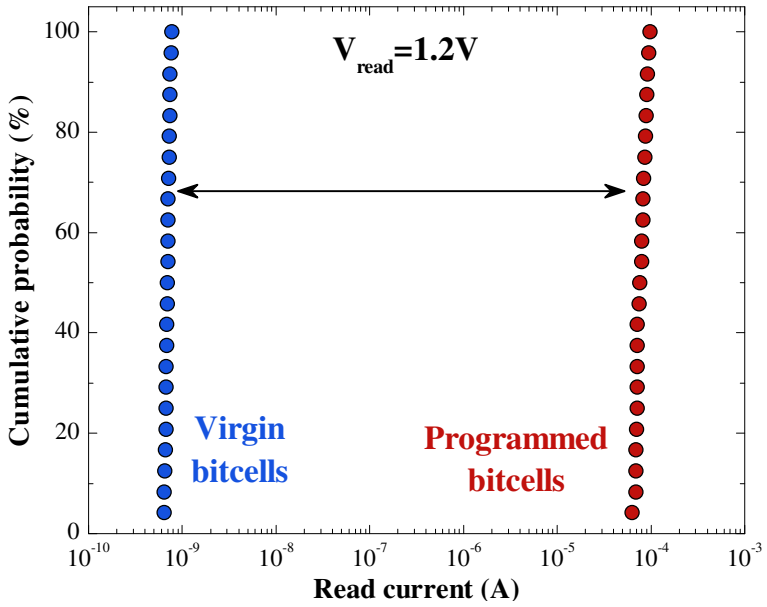
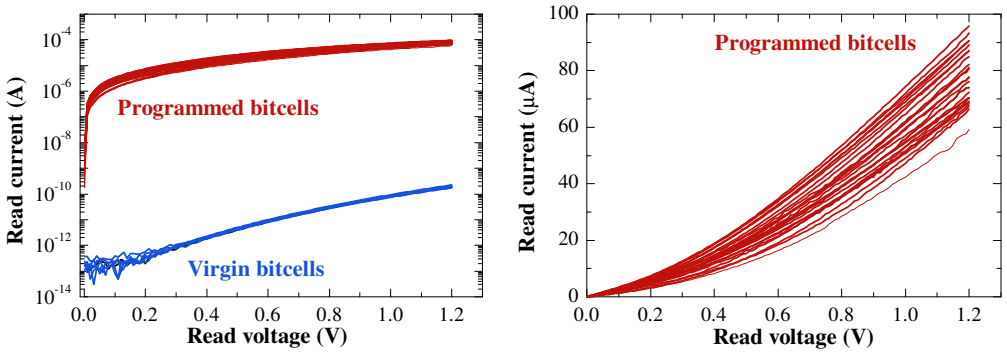


Figure 6.2: Read current distributions of virgin and programmed drift antifuse bitcells designed and fabricated in a logic 40-nm CMOS process.

The mean value of the narrow distribution of virgin bitcells is lower than 1nA. Programmed bitcells exhibit a read current drastically higher and the mean value is 75μA. In this example, the read current is significantly less dispersed than the time-to-breakdown (see an example in chapter 4, figure 4.11).
In an antifuse memory, bitcells are read using a current sensor. Therefore, the wider the difference in read current amplitude between a virgin and a programmed bitcell, the more effective the sensing circuit. However, the specification in read current amplitude must be counterbalanced with the power consumption during read operation.

6.1.2 Breakdown path characteristic

In the previous example, the read current was measured under a single voltage amplitude in order to distinguish a virgin and a programmed bitcell. This approach can be completed by measuring the leakage current over a voltage range and thus, plotting the electrical characteristic of the breakdown spot.
Example of characterizations are plotted in figure 6.3.



(a) (b)

Figure 6.3: $I_{read} - V_{read}$ characteristics of virgin and programmed drift antifuse bitcells plotted in a semilogarithmic scale (a) and linear scale (b).

The difference in read current amplitude between virgin and programmed bitcells is emphasized in figure 6.3(a). The linear plot shown in figure 6.3(b) gives interesting insights in the nature of the breakdown spot. Indeed, a programmed antifuse capacitor is not a mere resistive material. The current-voltage characteristic is not linear and exhibit a sort of exponential behavior. Due to this particular characteristic, it is difficult to read antifuse bitcells under a low voltage.

The effect of a severe heating of the breakdown spot is certainly involved in the degradation of the dielectric material. However, it is difficult to evaluate the local increase in temperature in such small devices. This aspect is therefore discarded in the first approach of this study.

The current-voltage characteristics appear linear between 0.8V and 1.2V. Therefore, measurement points were fitted using a linear trendline in order to identify the resistance value. Results are shown in figure 6.4.

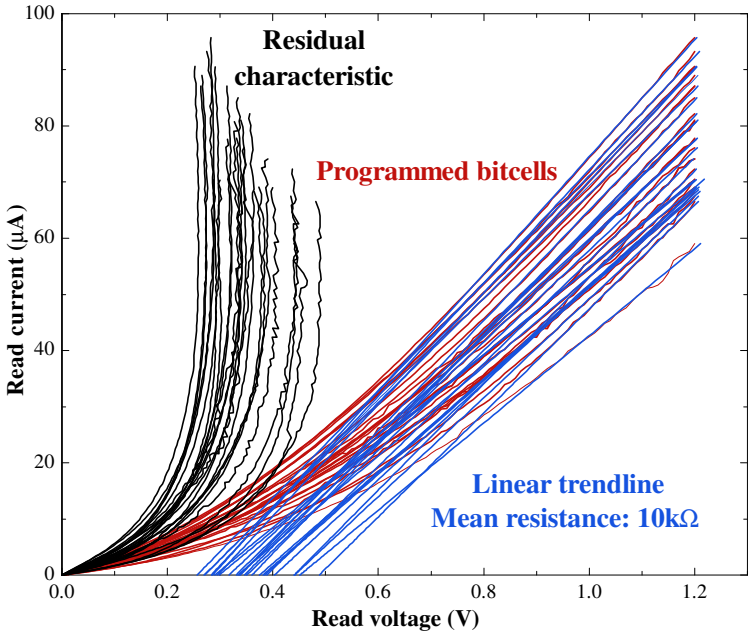


Figure 6.4: Identification of the resistance value R_{prog} from current-voltage characteristics of programmed antifuse bitcells using linear trendlines. The residual characteristics are calculated as $V_{residual} = V_{read} - R_{prog} \cdot I_{read}$.

Linear trendlines are correctly fitted against measurements. A mean resistance of 10kΩ was identified. Note that the on-state resistance of the access transistor is taken into account.

A residual characteristic was calculated as $V_{residual} = V_{read} - R_{prog} \cdot I_{read}$. In other words, it is assumed that the resistor identified using a linear trendline is connected in series with another device exhibiting the so-called residual characteristic.

Resistance values between 9.5 and 13kΩ were identified from linear trendline. The dispersion is therefore not significant. Exponential interpolations of residual characteristics were attempted without any success.

6.1.3 Perspectives

The identification of a significant resistance in the breakdown spot characteristic allows a modeling approach of a programmed antifuse bitcell as depicted in figure 6.5.

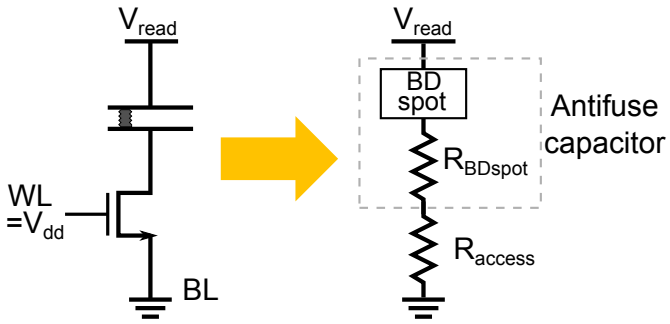


Figure 6.5: Proposed equivalent circuit of a programmed antifuse bitcell.

A possible equivalent circuit of a programmed antifuse bitcell comprises the breakdown spot connected in series with two resistors such as the on-state resistance of the access transistor and the resistive part of the breakdown spot. This approach is relevant if the characterization voltage is in a range such that the breakdown spot is not altered.

Some models of post-breakdown gate-current are reported in literature. The equivalent circuits of the breakdown path are more complex than the single branch depicted in figure 6.5. Nevertheless, models combine series element and diodes or Schottky barriers to deal with the exponential behavior [112–115].

The modeling of the residual characteristic using a trap-assisted current transport process such as Frenkel-Poole is also a possibility (see in chapter 3 section, 3.2.1.3). Indeed, defects are generated in the gate-oxide during the wearout phase.

This first study of the broken gate-oxide provided a better insight into the electrical characteristic of a programmed antifuse bitcell. However, guidelines or sweet spot relevant to reduce to the energy consumed in the post-breakdown phase cannot be pointed out using the present model. In this chapter, the contribution of the post-breakdown time and the post-breakdown current amplitude on the breakdown path characteristic is studied experimentally by comparing the effect of the programming conditions on the read current.

6.2 Impact of the post-breakdown conditions

The read current amplitude is key parameter in the performance of an antifuse memory. There is therefore an interest in the understanding of the impact of programming conditions on the resulting read current. As mentioned previously, the post-breakdown phase is defined by two main parameters such as the post-breakdown time and the post-breakdown current amplitude. Hence, it is worth to study the influence of one parameter or the other on the post-breakdown path characteristic.

The post-breakdown time depends on the time-to-breakdown and the width of the programming pulse. Consequently, it cannot be accurately set on a standard antifuse bitcell. Assuming that the programming time is significantly longer than the time-to-breakdown, the post-breakdown time is set by the programming pulse width. This is a likely condition since minimum programming time of 10μs are reported whereas time-to-breakdown shorter than 100ns were measured (see in chapter 4).
The post-breakdown current amplitude is also uncontrolled. Chapter 5 introduced a side effect during programming which leads to a severe overshoot on the bulk current. The current flowing through the breakdown spot is therefore the combination of the bitline and bulk current.
The study of the post-breakdown phase requires particular test structures in order to control the post-breakdown time and current amplitude. In this section, the contributions of the post-breakdown current amplitude on the read current and the breakdown path characteristic are investigated. The programming time is set such that the time-to-breakdown can be neglected.

6.2.1 Post-breakdown current limiter circuit

The post-breakdown current is not limited by the access transistor due to the bulk current overshoot presented in chapter 5. A solution was implemented in the test structure in order to accurately control current flowing through the breakdown path.

6.2.1.1 Topology and design

Since the access transistor does not achieve current limitation, another device is needed between the HV node and the gate of the antifuse capacitor in order to control the current flowing through the broken dielectric material. An example of test structure schematic is depicted in figure 6.6.

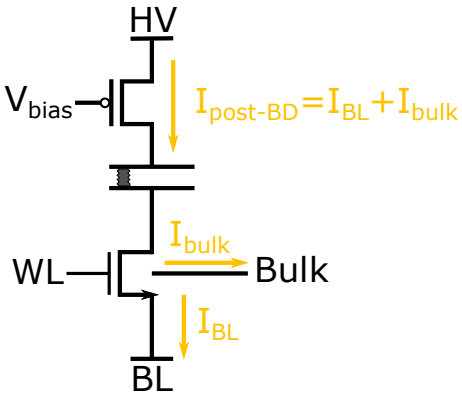


Figure 6.6: Example of test structure enabling the control of the post-breakdown current amplitude.

The post-breakdown current is limited by the top PMOS transistor. The amplitude is set by the size of the latter device and the bias voltage V_{bias}. The bulk current overshoot is not disabled by this structure. However, the fact whether the current goes to the bitline or the bulk node is irrelevant as the role of this structure is to control the overall post-breakdown current amplitude.
The specifications for the appropriate design of this structure are listed below.

- **Low on-state resistance:** Since a series device is connected to the bitcell, the drain-to-source voltage must be minimized during the wearout phase. A maximum drop of 100mV is tolerated.
- **Low drain capacitance:** After the breakdown event, the drain voltage of the PMOS transistor changes in order to set the post-breakdown current amplitude. The settling time depends on the latter amplitude and the drain capacitance. It should not exceed 100ns.
- **Tunable post-breakdown current:** This feature is required to study the contribution of the post-breakdown current amplitude on the breakdown path characteristic. A range from 20μ to 800μA is specified for the investigations.
- **High voltage operation:** The voltage applied across the PMOS transistor can be quite high during the post-breakdown phase. The safe operation must be insured, i.e. in a nominal voltage range.

An equation of the on-state resistance of a MOS transistor as a function of the gate-to-source voltage and the saturation current is derived in equation 6.1.

$$R_{on} = \frac{V_{gs} - V_t}{2 \cdot I_{dsat}} \tag{6.1}$$

I_{dsat} is set according to the specification of the post-current amplitude. Thus, R_{on} can be minimized for a low V_{gs}, i.e. a large transistor width. However, this trend leads to a larger drain capacitance.
A schematic of the post-breakdown current source designed and fabricated in a logic 40-nm CMOS process is depicted in figure 6.7. Either a drift or cascode bitcell can be implemented.

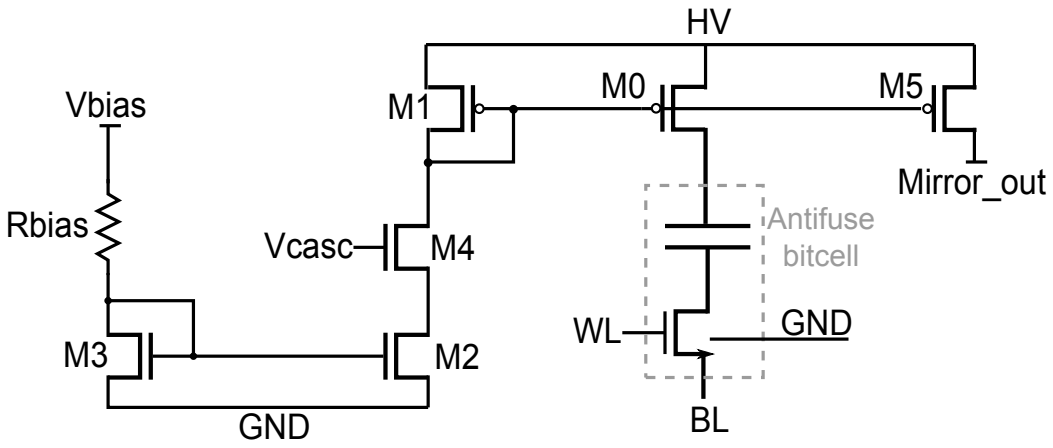


Figure 6.7: Schematic of the post-breakdown current source as designed and fabricated in a logic 40-nm CMOS process.

The post-breakdown current amplitude is set by `Vbias` in the left branch of the current mirror formed by `M3` and `M2`. Then, this current is copied in another current mirror formed by `M0`, `M1` and `M5`. The device `M4` is a cascode transistor which prevents `M2` to operate in a high voltage range. `M5` is used to characterize the current source prior to experiments.
The current source was designed using thick-oxide devices (50Å). The nominal supply voltage range of 2.5V is sufficient to perform programming operation under a voltage up to 7V.
This circuit was embedded in a test chip with the drift antifuse bitcells listed in chapter 4, table 4.5. Experiments were performed on the small and medium bitcells.

6.2.1.2 Performance

Electrical characterization were performed on the circuit prior to experiments. The reference current range set using `Vbias` and the output current (`Mirror_out`) were measured. Characteristics are plotted in figure 6.8.

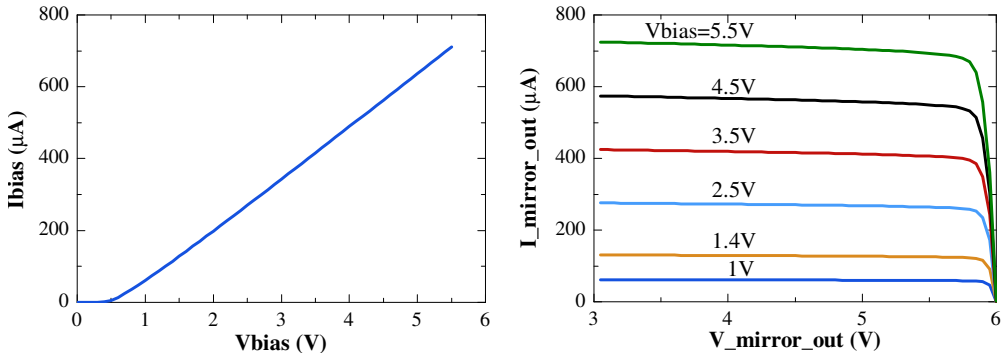


Figure 6.8: Ibias-Vbias characteristic of the current source (a). Output current measured on `Mirror_out` for various `Vbias` amplitude (b).

The reference current of the source can be set from 25μA to 750μA. The measurements of the output current for various `Vbias` show that the reference current is accurately copied over a wide voltage range.

The transistor `M0` was dimensioned in order to feature an on-state resistance of 200Ω.

An example of a practical experiment is shown in figure 6.9. Waveforms of I_bitline and I_bulk were measured during a programming operation of 50μs under 5.5V.

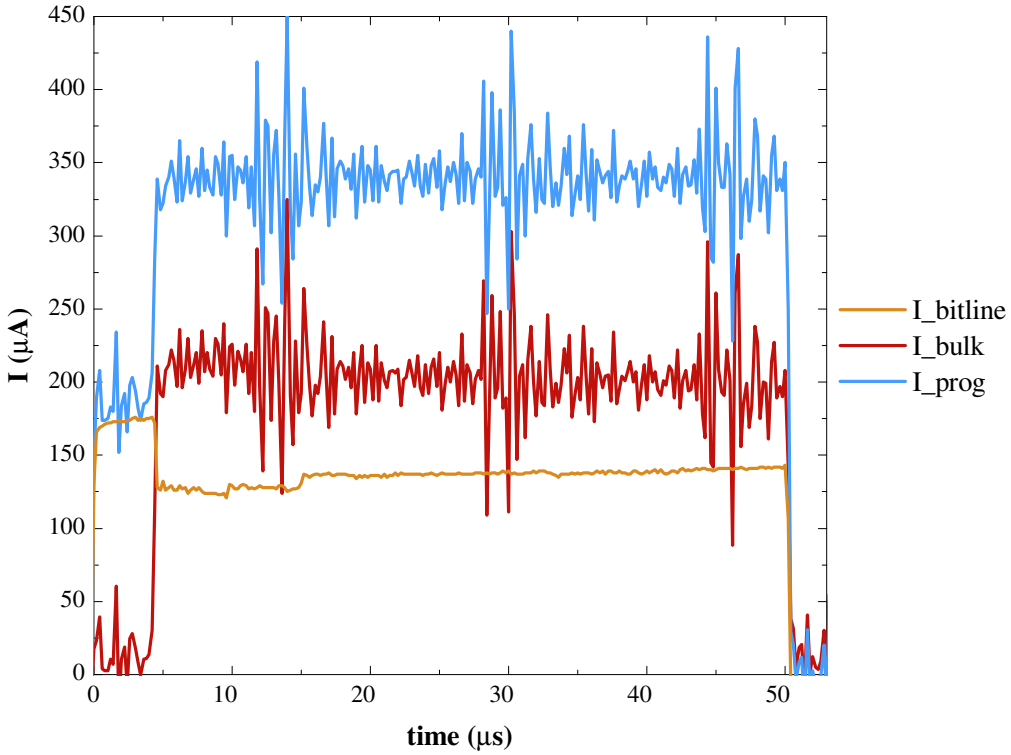


Figure 6.9: Waveforms of the bitline and bulk current during a programming operation. I_prog is calculated as I_prog=I_bitline+I_bulk.

During the wearout phase, the bitline current reaches 175μA for approximately 5μs. Then, the breakdown event occurs. The post-breakdown current reaches an average amplitude slightly lower than 350μA as set by `Vbias`. The bulk current waveform is noisy (so is I_prog) due to the configuration of the prober [1]. Although this noise is annoying for the visualization of the bulk current waveform, it does not affect the current flowing through the breakdown spot.

Since the post-breakdown current is limited by the current source, its amplitude is divided between the bitline and bulk current. As a matter, it is seen that the bitline current becomes lower in the post-breakdown phase than during wearout. The bulk reaches a certain value which comply with the total post-breakdown current amplitude of 350μA.

The bitline and bulk currents vary along the post-breakdown phase. I_bitline increases slightly between 15 and 50 μs whereas I_bulk decreases with the same pace. Therefore, I_prog is stable at the amplitude set by `Vbias`. As observed in chapter 5, the mechanism leading to the bulk current overshoot seems attenuated. This variation is compensated by the bitline current.

[1] The same noise signal was measured by connecting an oscilloscope on the prober chuck without applying any electrical signals and without a wafer

The proper operation of the current source is validated in this example. The post-breakdown current amplitude is constant during the post-breakdown phase and reaches the level set by `Vbias`.

6.2.2 Read current distributions & characteristics

6.2.2.1 Read current distributions

Programming operations were performed on drift antifuse bitcells under different conditions of post-breakdown current amplitude and programming time. Read current distributions are plotted in figure 6.10. The programming conditions are detailed in legend. For comparison purpose, the distributions of a standard virgin and programmed bitcell previously shown in figure 6.2 are displayed in the figure below (the post-breakdown current is not limited).

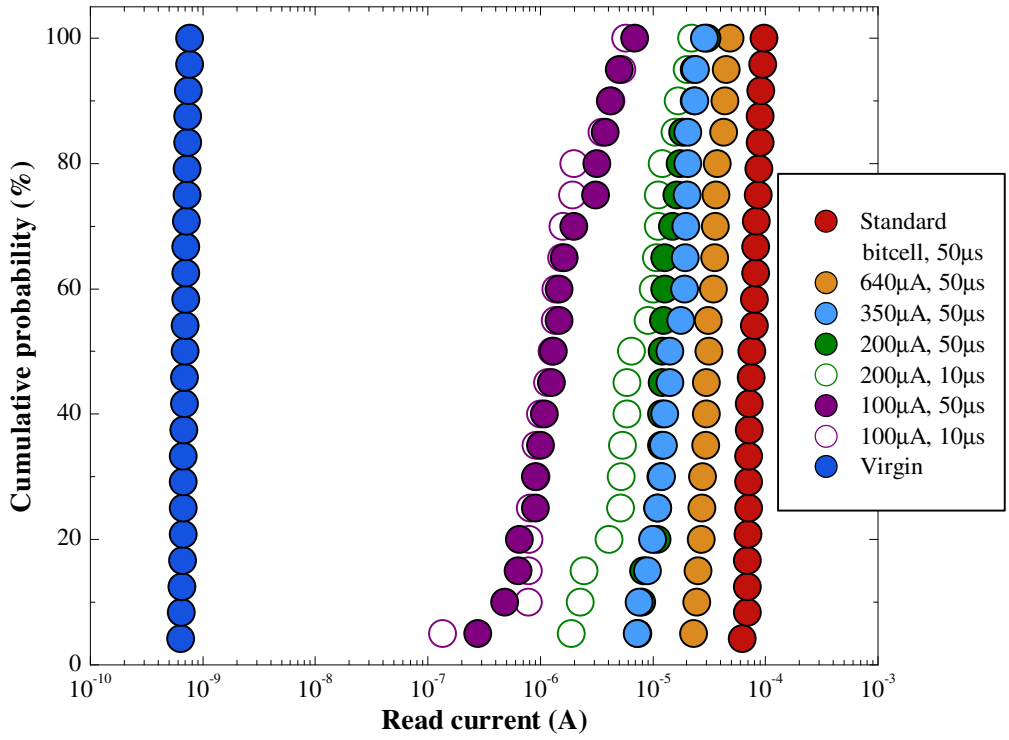


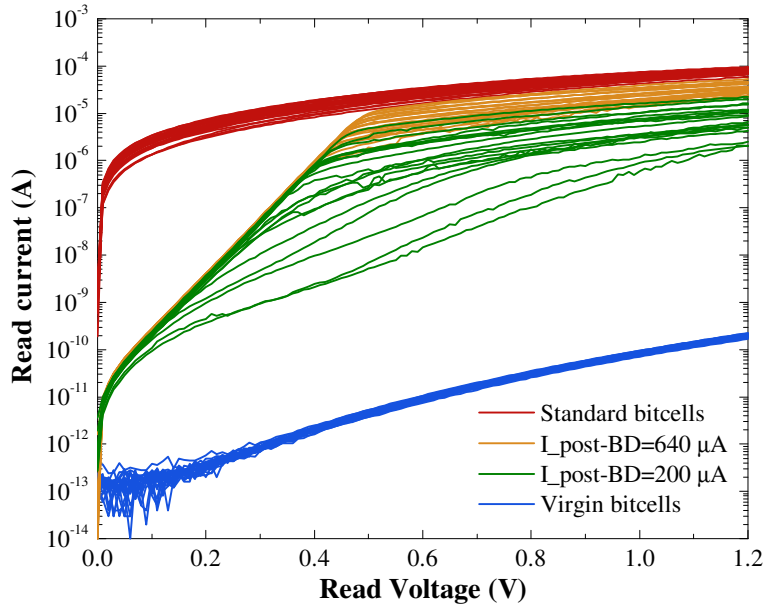
Figure 6.10: Read current distributions of drift antifuse bitcells designed and fabricated in a logic 40-nm CMOS process. Programming operations were performed with a constant post-breakdown current under HV=5.5V.

The mean value of the read current distributions is impacted by the post-breakdown current amplitude. A trend is emphasized as the lower the post-breakdown current, the lower the read current. The dispersion of the distribution is also affected. The dispersion is smaller for the standard bitcells and the devices programmed with a

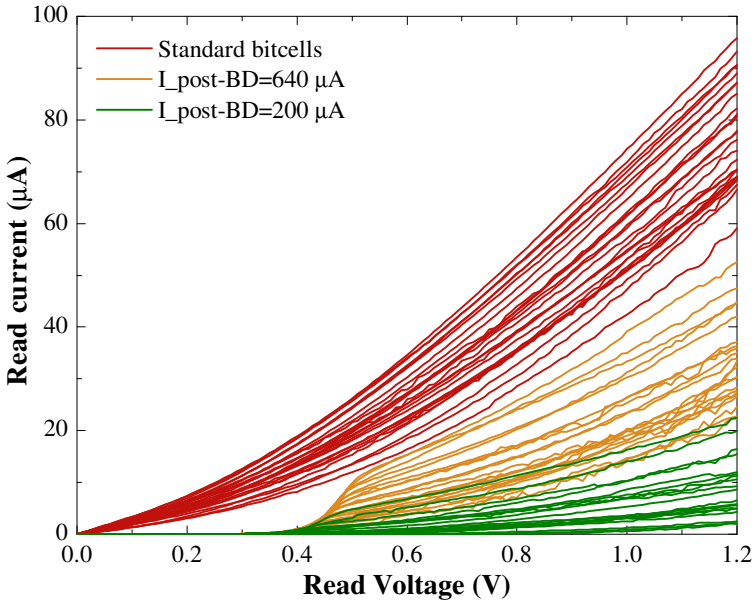
current amplitude of 640μA whereas for a current of 100μA, read current values are spread over more than one order of magnitude.
Although the post-breakdown time cannot be accurately controlled using this test structure, programming operations were performed for shorter time, e.g. 10μs, the post-breakdown current was set either at 100 or 200μA. The programming time appeared to have an effect on the read current distribution corresponding to a post-breakdown current of 200μA. The dispersion is indeed narrower for a programming time of 50μs. However, the distributions are similar for a post-breakdown current of 100μA.
There is also a contribution of the post-breakdown time on the read current. However, it seems bonded with the carriers energy that damages the dielectric material. A change in the post-breakdown current leads to a different effective voltage applied to the bitcell due to the saturation of the transistor `MO`. As a consequence, setting a low post-breakdown current leads to less carriers with less energy that cannot create deeper defects in the dielectric material. There is probably a time range in which the read current can be increased until a maximum that depends on the carrier energy.
Further work is obviously needed to draw accurate conclusions on the contribution of the post-breakdown time on the read current amplitude. More particularly, a test structure enabling the accurate control of the post-breakdown time. A breakdown detection system is presented in chapter 7.

6.2.2.2 Breakdown path characteristic

DC characterizations were performed on antifuse bitcells corresponding to each read current value in the distributions plotted in figure 6.10. $I_{read} - V_{read}$ characteristics are plotted in figure 6.11 only for a post-breakdown current of 640 and 200μA for the sake of clarity. Curves for standard and virgin bitcells are recalled from figure 6.3 for comparison purpose.



(a)



(b)

Figure 6.11: $I_{read} - V_{read}$ characteristics of virgin and programmed drift antifuse bitcells plotted in a semilogarithmic scale (a) and linear scale (b).

Despite the subthreshold conduction of `M0` from 0 to 0.5V, the impact of the post-breakdown current amplitude on the characteristic of the breakdown path is clearly emphasized. The difference in read current amplitude is due to a difference in the resistance of the breakdown spot. The so-called residual characteristic may be also affected. Unfortunately no conclusion can be drawn on the latter point due to the

subthreshold conduction of M0.

The resistance of different programmed bitcell was identified using a liner trendline as shown previously in figure 6.4. However, the residual characteristic is not studied. An example is shown in figure 6.12 for the bitcells programmed for 50μs and a post-breakdown current of 100μA.

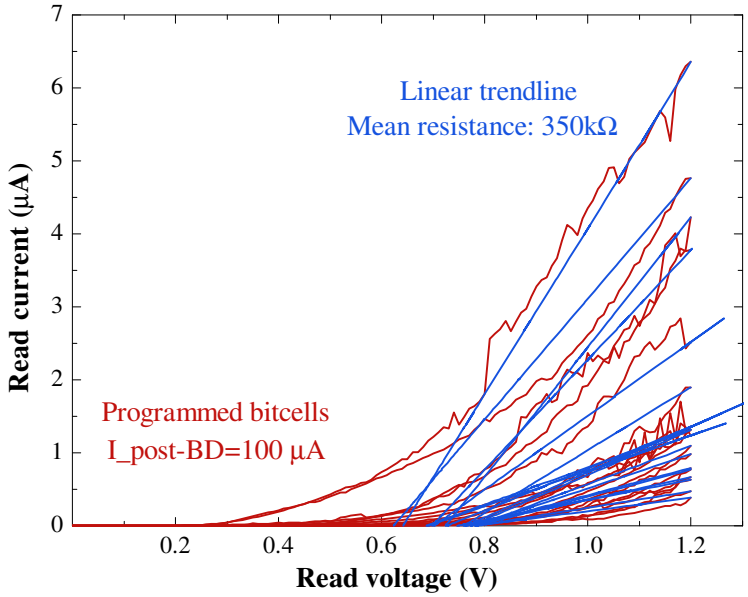


Figure 6.12: $I_{read} - V_{read}$ characteristics of drift antifuse bitcells programmed under a voltage of 5.5V, for 10μs and for a post-breakdown current of 100μA. The resistance of each bitcell was identified using a linear trendline.

The dispersion of the read current is reflected by the different trendlines. This observation confirms that the resistance of the breakdown path is impacted by the post-breakdown current. The mean resistance identified from trendlines is 350kΩ whereas a value of 10kΩ was characterized on standard bitcells, i.e. without post-breakdown current limitation. However, the wider dispersion is not explained. The characteristic of the breakdown spot is particular. Indeed the curves appear unstable whereas conventional DC characterizations should yield smooth lines. The reason of this particular signature may be that the gate-oxide is not completely broken down due to the low post-breakdown current. As a consequence, charge trapping may occur and yields unstable characteristics. Even though there is a difference of three orders of magnitude between virgin bitcells and devices programmed with a post-breakdown current of 100μA, the state of the gate-oxide is somehow between a insulating and a burnt-out dielectric material. In other words, the energy of carriers injected through the dielectric material is not sufficient to obtain a hard breakdown state.

Figure 6.13 concludes this discussion. Distributions of resistance values identified from prior characterization are plotted.

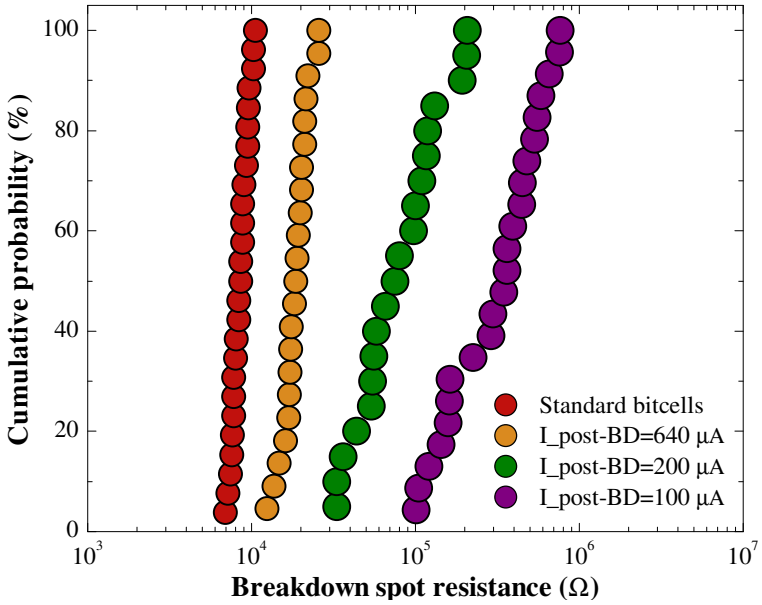


Figure 6.13: Distributions of resistance values identified from $I_{read} - V_{read}$ characteristics. Programming operations were performed with a constant post-breakdown current.

The resistance follows the same trend of the read current: the higher the post-breakdown current, the lower the resistance, the higher the read current. The dispersion is also similar. Indeed, the distribution for the bitcells programmed with a post-breakdown current amplitude of 100μA exhibits a dispersion of roughly one order of magnitude whereas the dispersion of resistance values for standard bitcells is much smaller.

The results from the experiments reported in this section emphasized the role of a series resistor in the breakdown path. This modeling approach exhibits relevant descriptive properties. However, it cannot be used for a design purpose due to the approximations and the dispersions in the breakdown spot characteristics.

6.2.2.3 Cascode bitcell

Similar experiments were performed on cascode bitcells designed and fabricated in a logic 40-nm CMOS process. This bitcell architecture was characterized in chapter 4, section 4.5. Dimensions are listed in table 4.7.

Cascode bitcells were programmed by a voltage pulse of 6.6V amplitude and durations of 10μs or 1μs. Since the mean time-to-breakdown characterized for this bitcell architecture is 6ns, the programming time can be drastically reduced for

these studies. The wearout current is also much lower than in a drift antifuse bitcell.

Read current distributions are plotted in figure 6.14 for different programming conditions.

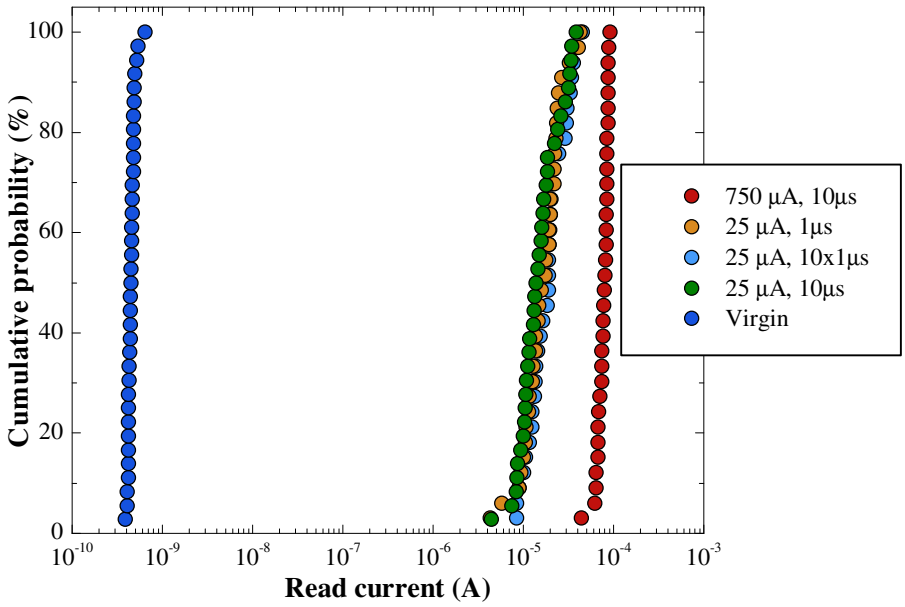


Figure 6.14: Read current distributions of cascode antifuse bitcells designed and fabricated in a logic 40-nm CMOS process. Programming operations were performed with a constant post-breakdown current.

The bitcells programmed with a high post-breakdown current amplitude exhibit a read current of about 100μA and a small dispersion.

Programming operations were performed with a very low post-breakdown current of 25μA and a programming time of 10μs as a first approach. The read current distribution exhibits a mean value of 147μA. Like for drift antifuse bitcell, a lower read current is obtained for a lower post-breakdown current. However, the value is significantly higher than the mean read current measured on a drift antifuse bitcell programmed under 5.5V and a post-breakdown current of 100μA. Since the test structure is the same, the difference in read current could be explained by the higher effective programming voltage applied to the capacitor.

Different programming time were tested on the cascode bitcell. The distribution obtained for $T_{prog} = 1\mu$s is similar to the one corresponding to a programming time of 10μs. In other words, it seems that maintaining a high voltage for more than 1μs has no effect on the breakdown spot. This assumption is confirmed by studying the evolution of the read current distribution after successive pulses of 1μs. Indeed, the mean value and the dispersion of the distribution are similar even after 10 successive programming sequences.

Waveforms of the bitline current and the wordline voltage are given in figure 6.15. The bulk current is not displayed due to noise issues on the prober.

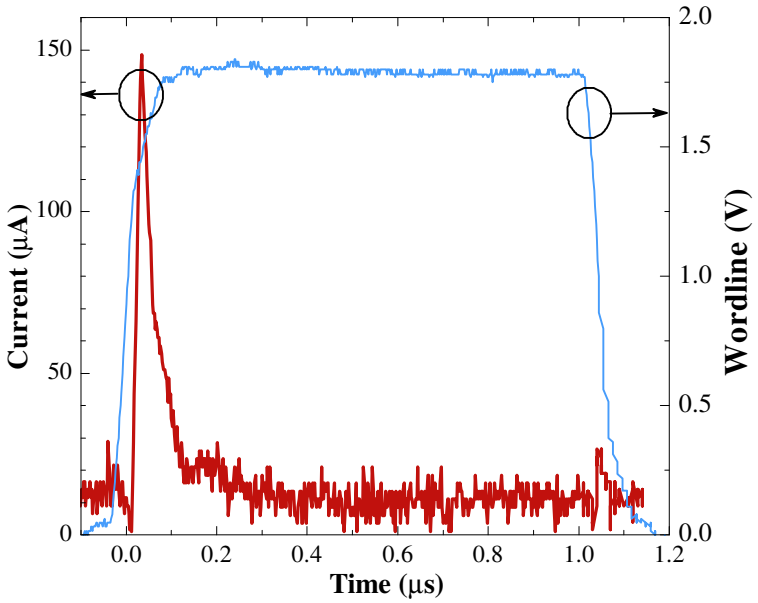
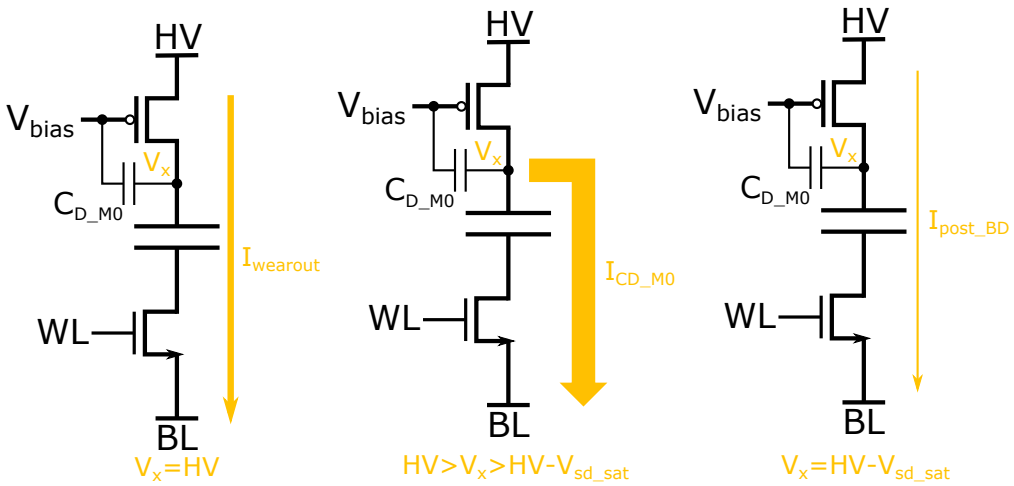


Figure 6.15: Waveforms of the bitline current during a programming operation. The programming pulse is applied to the wordline to turn on the access transistor for 1μs.

The bitline current was measurement using a 50-Ω series resistor. The setup is detailed in chapter 3, section 3.4.2.3. Although the bandwidth is limited and does to allow to track a transient time down to 1μs, the accuracy is sufficient to investigate qualitatively the short programming sequence.
In fact, the programming sequence can be divided in three steps as depicted in figure 6.16.



(a) Wearout (b) Breakdown (c) Post-Breakdown

Figure 6.16: Illustration of a programming phase in wearout (a), breakdown (b) and post-breakdown (c) phases.

1. The wearout current is significantly higher than the post-breakdown current set by the current source. However, the voltage applied across the bitcell is not affected during the very short wearout phase.

   The drain capacitance is depicted in figure 6.16(a). In fact, this parasitic capacitor decouples the node V_x from `M0`. As a consequence, the programming voltage of 6.6V is maintained across the bitcell during the wearout phase.

2. When the gate-oxide breakdown occurs, the capacitor is discharged through the bitcell as shown in figure 6.16(b). This the reason why a peak current up to 150μA is seen in figure 6.15. This discharge takes approximately 200ns. During this transient phase, the voltage at the node V_x decays from HV to $HV - V_{sd\_sat\_M0}$.

3. A steady state is reached once `M0` is saturated. Since the post-breakdown current is set at a quite low level in this experiment, a severe voltage drop can be expected across the current source. As a consequence, the number of carriers damaging the dielectric material is low and so is their energy.

The breakdown event occurs after roughly ten nanoseconds as observed in chapter 4. The bitline current reaches 150μA and decreases down to 20μA due to the current source set at 25μA. The effect of the drain capacitance of `M0` is clearly seen in this example. A settling time of 200ns is needed to limit the current down to the reference level. After the peak, the bitline current remains stable.

Since a programming time of 1μs or 10μs has the same effect on the resulting read current. It can be assumed that the period during which the current is limited by the current source has no effect on the damage of the dielectric. Hence, the breakdown spot characteristic is set during the transient of 200ns.
During the transient event occurring right after the breakdown event, the drain capacitor of `M0` has to be discharged through the breakdown path. As long as the current flows from the capacitor to the bitline, the drain voltage of `M0` decreases until the operating point set by the reference current. During this short transient event, highly-energized carriers flow through the gate-oxide and may cause severe damages.
The use of a post-breakdown current limiter circuit is relevant to avoid a high power consumption due to the bulk current overshoot (see in chapter 5). The contribution of the programming current on the read current amplitude can be therefore analyzed a main trend was emphasized as the higher the programming current, the higher the read current. However, the transient response of the current source has also a significant impact on the generation of defects in the gate-oxide. The reduction of the drain capacitor of `M0` and therefore the settling time is a room for improvement in order to control the post-breakdown current more accurately.

6.2.3 Discussion

The implementation of a current source on top of an antifuse bitcell is a relevant solution to overcome the high power consumption involved in the programming operation due to the combination of the bitline current and the bulk current overshoot. Since most of the programming time is spent in post-breakdown. Efforts to reduce the programming energy should be focused on this phase. A trend on the contribution of the post-breakdown current was highlighted in this section. However, the voltage may be also correlated to the degradation of the dielectric material. In other words, the number of carriers and their energy should be considered.
The post-breakdown current limiter circuit was proven relevant to study the role of the post-breakdown current amplitude in the breakdown path characteristic. However, this design can be improved especially by reducing the drain capacitance of the series PMOS transistor. An upgraded topology of current source is presented in chapter 7.

6.3 Conclusion

The last phase of a programming operation of an antifuse bitcell was studied in this chapter. The post-breakdown phase is defined by the post-breakdown current and

time. In other words, a certain amount of charge flows through the breakdown path after the gate-oxide breakdown that damages the dielectric material. As a result, a read current can be measured by applying a low voltage (1-1.5V) to the bitcell. It was shown in this chapter that the post-breakdown conditions have an impact on the read current. Since most of the programming energy is consumed in the post-breakdown phase as a result of a high current and a long time, the objective of the present study is to understand how to minimize the programming energy during this phase while obtaining an acceptable read current amplitude.

A programmed bitcell can be characterized either by measuring the read current and by plotting the cumulative distribution, or by applying a voltage ramp on the gate of the antifuse capacitor and plotting the characteristic $I_{read} - V_{read}$. Experiments have shown that the latter characteristic involves a resistance connected in series to a device exhibiting a sort of exponential characteristic like a diode or a schottky barrier. The objective of this chapter was to understand the contribution of the post-breakdown phase on the amplitude of the read current.

A dedicated test structure was designed in order to control the amplitude of the post-breakdown current. Indeed, the control of the programming current is difficult due to the bulk current overshoot mechanism. Read current distributions were plotted for different post-breakdown current conditions. As a result, the sets of bitcells programmed with a lower post-breakdown current exhibited wider distributions with a lower mean read current value. $I_{read} - V_{read}$ characteristics emphasized that the lower the post-breakdown current, the higher the resistance.

Experiments were also performed on cascode antifuse bitcells. Due to the lower wearout current and shorter time-to-breakdown, programming operations were performed under a higher voltage and a very low post-breakdown current (25μA). An interesting result came out of this study. A mean read current of 10μA was obtained with a programming time of 1μs and 10μs. For both conditions, distributions exhibited a small dispersion and the same mean value. The reason why the programming time was not relevant on the definition of the read current comes from the fact that a high transient current flows through the bitcell due to the discharge of a parasitic capacitor from the current source circuit. During this period, the current is not yet limited by the circuit and the voltage applied to the bitcell is very high. Once the current compliance is set, this voltage is lowered and the dielectric no longer damaged. Finally, the read current is onset only during the transient event.

This result is promising for the validity of low energy programming. Even though this study remains in a early stage, it was shown that an acceptable read current can be obtained by programming antifuse bitcells in a very short time and a low current. Implementing a current source is therefore relevant to limit the post-breakdown

current, thereby the programming energy.
Many questions remains unanswered about the control of the read current. Nonetheless, it can be concluded that the state of the breakdown spot is onset during the post-breakdown phase. The generation of defects and their effect on the damage on the dielectric material depends on the energy of the injected carriers. Low-energized carriers cannot sufficiently wear the gate-oxide and yield a high read current even for a long post-breakdown time. On the other hand, a higher programming current leads to a higher read current due to the contribution of high-energized carriers. The resulting read current is therefore higher.
Another test structure is needed in order to accurately control the time between the breakdown event and the end of the programming sequence. This type of solution is easier to implement in a memory system because digital electronics is helpful to control a time. In addition, a higher memory density can be studied compared to a single test structure. The design of demonstrator is detailed in the next chapter.

CHAPTER 7

32-nm CMOS Advanced antifuse memory demonstrator

The study of the gate-oxide breakdown physics detailed in the previous chapters brought relevant knowledge. Observations, electrical characterizations and models emphasized that the duration of the wearout phase is significantly shorter than the programming time reported in literature (see in chapter 2, table 2.6). Consequently, most of the programming sequence is spent in the post-breakdown phase in which the current amplitude is the highest. Moreover, the bulk current overshoot triggered by the breakdown event leads to an uncontrolled energy consumption.

Considering the latter observations, the efforts to control the programming energy should be focused on the post-breakdown phase. Preliminary works are reported in chapter 6 and emphasized the role of the programming current amplitude on the breakdown path characteristic. A test structure comprising a current source and an antifuse bitcell was designed in order to control the post-breakdown amplitude. This feature is necessary to avoid the effect of the bulk current overshoot on the breakdown spot and therefore on the resulting read current amplitude.

The design of an advanced antifuse memory demonstrator is reported in this chapter implementing improvements issued from the results reported in the previous chapters. The two main goals are the implementation of an advanced programming modes that enables the control of the post-breakdown current amplitude and duration and the study of read current dependence on the post-breakdown conditions. The results reported in chapter 6 were obtained on test structures comprising a single antifuse bitcell. The environment of such a device in a memory system is different. The functionality and the benefit of an advanced programming mode

must be therefore validated on a memory system that comprises a dense array, as implemented in a product.
The operating modes and the specifications of the 1-kb demonstrator are detailed in section 7.1.
The design and the topology of the current source that controls the post-breakdown current is presented in section 7.2.
A programming detection system is introduced in section 7.3.
Conclusions are drawn in section 7.4.

7.1 Key features and operating modes

The specifications of the antifuse memory demonstrator are detailed in the present section. The prime objective is to implement and validate innovative features such as a programming current limiter and a programming detection system in a memory system.

7.1.1 Functionalities

A block diagram of the single antifuse memory demonstrator is depicted in figure 7.1. This circuit is designed and fabricated in a logic 32-nm CMOS process.

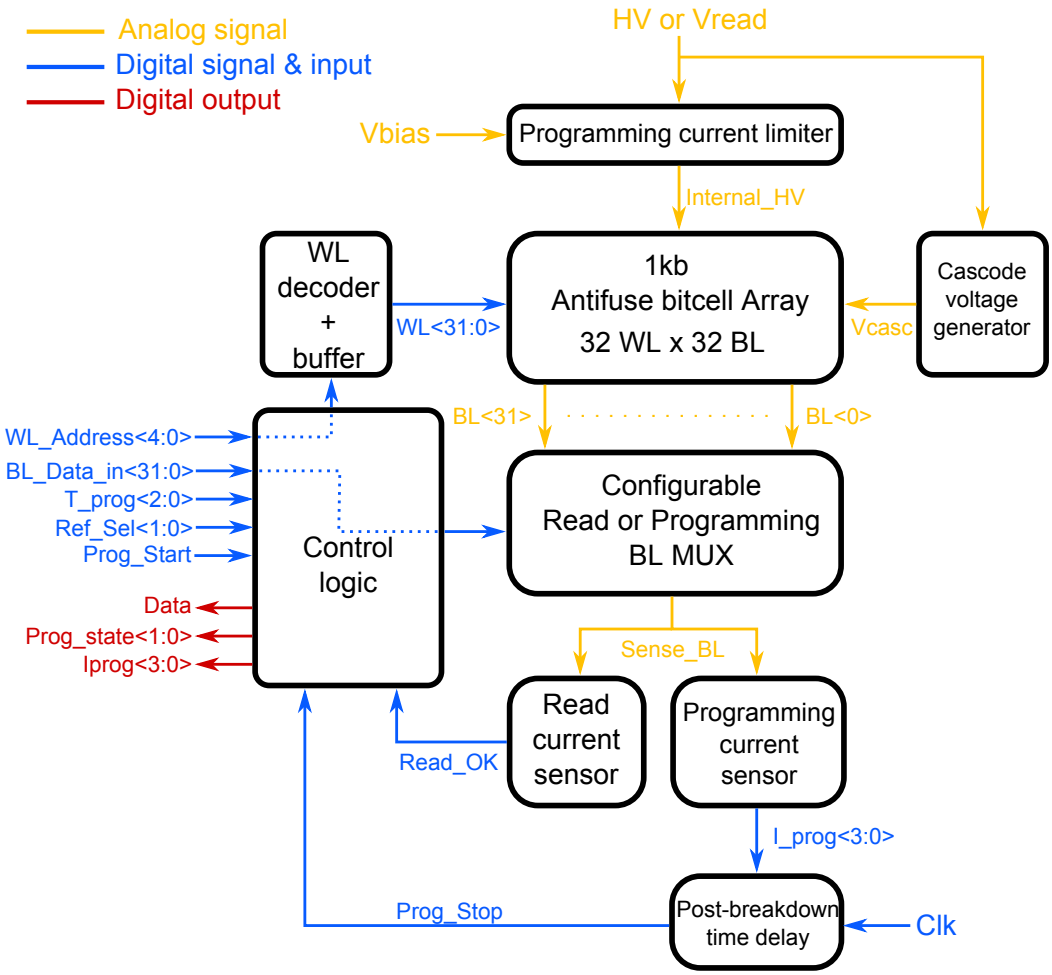


Figure 7.1: Block diagram of the antifuse memory demonstrator.

The central circuit block is the memory array. Cascode bitcells are organized in an array of 32 words of 32 bits. Information about the cascode bitcell designed in a logic 32-nm CMOS process are available in chapter 4, section 4.5.
The memory array is surrounded by different circuits. The system is operated using a control logic that sets three operating modes such as an advanced programming scheme, the standard programming scheme and a read operation. The supply voltages [1] including HV and Vread are also set by the user.

7.1.1.1 Standard programming mode

In the standard programming mode, a bitcell is selected and then, programmed by a high voltage pulse of 10μs typically. The memory array is addressed by the

[1] The control logic is designed using thin-oxide transistors supplied by `Vdd=1.1V`. The other circuits are designed using thick-oxide devices and are supplied by `Vdd1=1.8V`

signals WL_Address<4:0> and BL_Data_in<31:0> that access a selected wordline and bitline respectively. The selection of a bitline is achieved using a multiplexer either configured in a programming or read mode. Details on the topology of this circuit are given further in the chapter.
The standard programming mode is illustrated in figure 7.2.

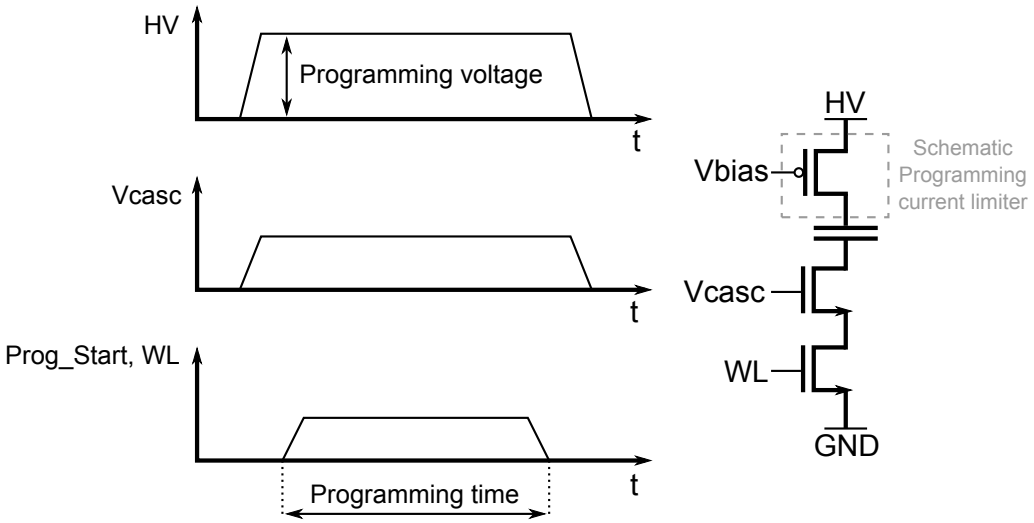


Figure 7.2: Timing diagram of the standard programming mode for a cascode antifuse bitcell.

The high voltage rises up to the programming voltage amplitude (typically 5.5V in 32-nm CMOS). The cascode voltage Vcasc is generated using a voltage divider. Since the bitcell is not selected, the antifuse capacitor is not stressed. The programming operation is performed by applying a pulse on Prog_Start. As a consequence, the control programming pulse is applied on WL. The access transistor is turned on, thereby stressing the dielectric material.
The programming current limiter is initially set before the rising of HV, eventually in a highly active state.

7.1.1.2 Advanced programming mode

The addressing method, the setting of the programming current limiter and the generation of the cascode voltage are the same as in the standard mode. However, the post-breakdown time is set by a detection system that comprises the circuits Programming current sensor and Post-breakdown time delay.
A timing diagram of an advanced programming operation is given in figure 7.3. Like in the standard mode, HV and Vcasc are set before starting the programming sequence.

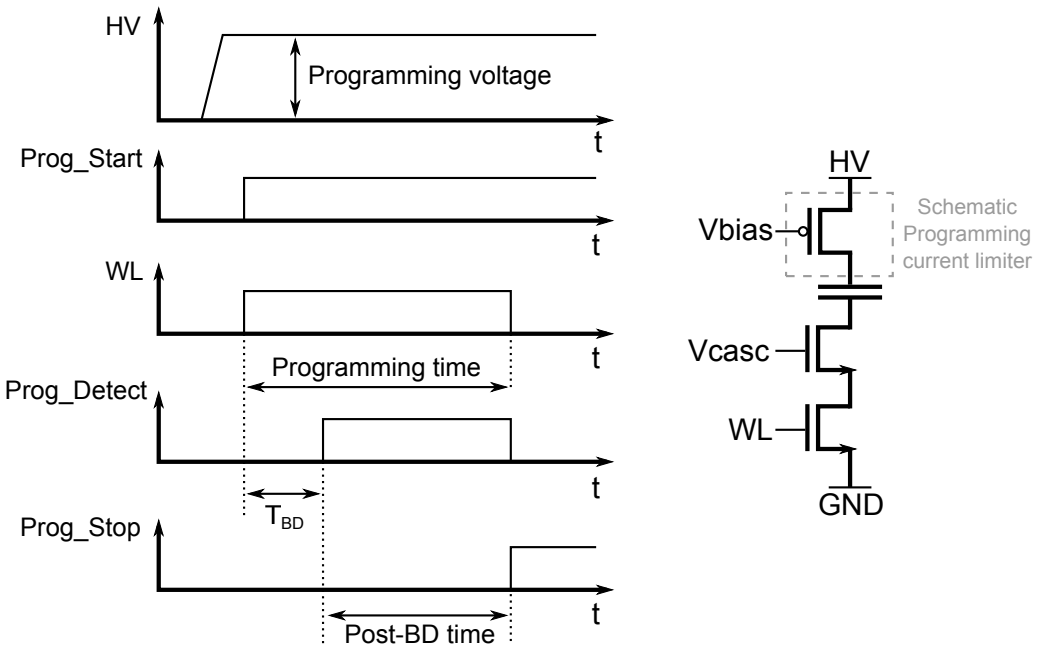


Figure 7.3: Timing diagram of the advanced programming mode.

Once `Prog_Start` rises up, so does `WL`. The antifuse capacitor is therefore being stress by the programming voltage. The programming current is sensed by `Programming current sensor` and compared to a reference current. The antifuse capacitor is considered broken when the leakage current becomes higher than the latter reference. This event is reflected by the rising of `Prog_Detect`. A time delay is then triggered in order to maintain the high voltage stress for a given time (Post-breakdown time). Once this time has elapsed, `WL` falls down to `0` and turns the access transistor off. The post-breakdown time and the detection current threshold are set by `T_Prog<2:0>` and by `Ref_Sel<1:0>` respectively. Operation ranges are presented in 7.1.2.
In the advanced programming mode, the post-breakdown current and time can be set independently. The objective of the present demonstrator is to investigate the influence of both parameters on the read current. Although, the impact of the post-breakdown current amplitude was discussed in chapter 6, the control of the post-breakdown time involves circuitry that is easier to implement in a complete system rather than in a single test structure. The programming detection system is further detailed in section 7.3.

7.1.1.3 Read mode

The purpose of the read mode is to measure the read current amplitude of programmed bitcells. The addressing is the same as either in a standard or advanced programming mode. The cascode voltage is set to the same amplitude as `WL`. The

programming current limiter circuit is disabled and the BL multiplexer is configured in read mode.
The bitcell current is compared to a configurable reference that is set by `Ref_Sel<1:0>`. The operating range is given in table 7.4.

7.1.2 Specifications

The specified performance of each operating mode is detailed in this section. The purpose of this demonstrator is to investigate the influence of different programming parameters such as the amplitude of HV, the post-breakdown current and the post-breakdown time. Consequently, the circuits are designed such that they feature a wide operating range as listed in table 7.1.

Signals	Min	Typ	Max
HV (V)	4	-	5.5
Vdd1 (V)	-	1.8	-
Vdd (V)	-	1.1	-
Iprog_Max (A)	100μ		1.3m
Iprog_threshold (A)	$64 \cdot$ Iref		$256 \cdot$ Iref
T_Prog (s)	300n		51.1μ
Vreads (V)	0.3	-	2.2
Iread_threshold (A)	$1 \cdot$ Iref		$20 \cdot$ Iref

Table 7.1: Specifications of the demonstrator in standard, advanced and read modes.

Assuming a `Clk=10MHz`, table 7.2 gives the post-breakdown time values corresponding to the configuration of `T_Prog <2:0>`.

`T_Prog <2:0>`	`000`	`001`	`010`	`011`	`100`	`101`	`110`	`111`
Post-BD time (s)	300n	700n	1.5μ	3.1μ	6.3μ	12.7μ	25.5μ	51.1μ

Table 7.2: Configurable post-breakdown time set by `T_Prog <2:0>`.

The threshold of the programming detection current is set by the signal `Ref_Sel <1:0>` and a reference current. Each configuration has a corresponding output signal `I_Prog<2:0>`.

Ref_Sel <1:0>	00	01	10	11
I_Prog<2:0>	I_Prog<0>	I_Prog<1>	I_Prog<2>	I_Prog<3>
Detection threshold	64.Iref	128.Iref	192.Iref	256.Iref

Table 7.3: Configurable programming current detection threshold set by `Ref_Sel <1:0>` and Iref. Each configuration has a corresponding output signal `I_Prog<2:0>`.

In a read mode, the read current threshold is also set by `Ref_Sel <1:0>` and a reference current.

Ref_Sel <1:0>	00	01	10	11
Read threshold	1.Iref	4.Iref	16.Iref	20.Iref

Table 7.4: Configurable read current detection threshold set by `Ref_Sel <1:0>` and Iref

7.2 Programming current limiter

The benefits of a programming current limiter were demonstrated in chapter 6. As a result, the post-breakdown current flowing through the breakdown spot is accurately controlled. An acceptable read current amplitude was obtained with a low post-breakdown current, thus the implementation of a programming current limiter in the demonstrator appeared valuable for validation. The contribution of the post-breakdown current can be studied on a large statistical population.
The design and the performances of the programming current limiter circuit are detailed in the present section.

7.2.1 Topology and design

The requirements for an appropriate programming current limiter are the same as presented in chapter 6. Since the drain capacitor of the PMOS transistor connected to the antifuse bitcell led to a post-breakdown current settling time of 200ns, the topology was modified in order to enhance the transient response of the circuit. A simplified schematic of the programming current limiter circuit is depicted in figure 7.4.

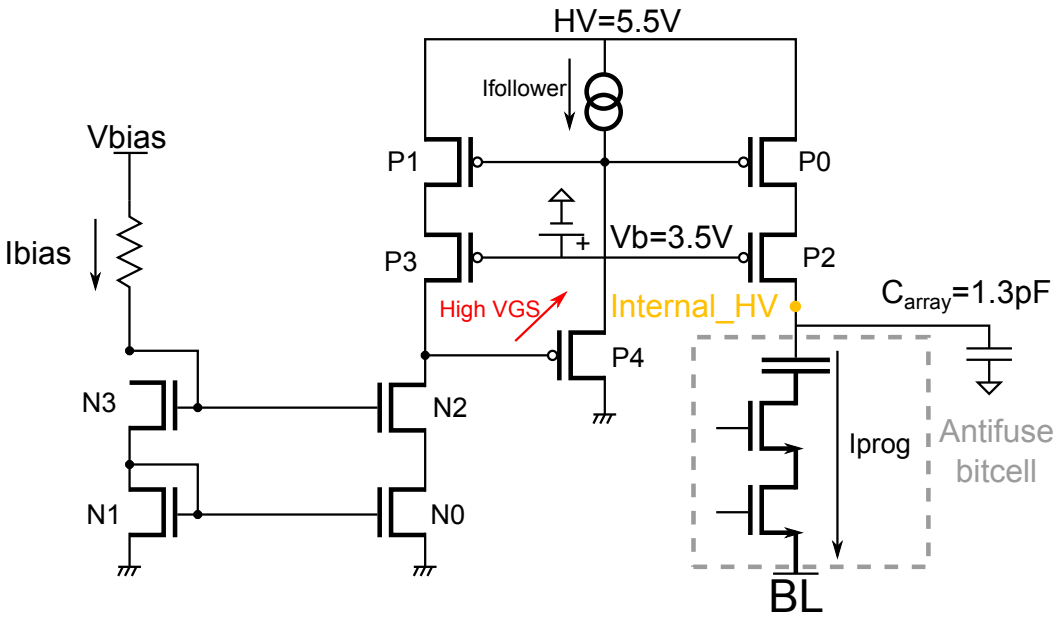


Figure 7.4: Simplified schematic of the programming current limiter implemented in the antifuse memory demonstrator.

The reference current `Ibias` is copied by the current mirror composed of `N0`, `N1`, `N2` and `N3`. A cascoded topology is used for accuracy. The bias current is also copied in the branch connected to the bitcell array (`P0` and `P2`).

The benefits of the cascode topology used in the PMOS stage are a lower output capacitance and a high voltage compatibility. `P0` and `P1` are dimensioned in order to feature a low on-state resistance as detailed in chapter 6, section 6.2. However, the cascode stage composed of `P2` and `P3` are driven by a voltage of 3.5V. Assuming a low voltage drop across `P0`, the gate-to-source voltage of `P2` is therefore high and its on-state resistance low. Consequently these two transistors can be implemented with a smaller width leading to drain capacitance five times smaller than in the test structure presented in chapter 6. The amplitude of `Vb` is chosen such that the drain-to-source voltage of `P0` and `P2` respectively remains in the nominal voltage range for any reference current and programming voltage amplitude.

A conventional cascode current mirror would be designed as depicted in figure 7.5. Due to the low amplitude of `Vb` and therefore the high gate-to-source voltage of `P2` and `P3`, a source follower stage is needed to insure the saturation of `P3`. Assuming a direct connection between the gate of `P1` and the drain of `P3`, the drain-to-source voltage of `P3` would not be sufficiently high to saturate the latter transistor. Consequently, the current copy would not be accurate. The implementation of a source follower stage composed of a current source and `P4` enables to lower the drain volt-

age of P3. This is the reason why P4 is dimensioned such that its gate-to-source voltage is sufficiently high to saturate P3 for any configuration of the current limiter.

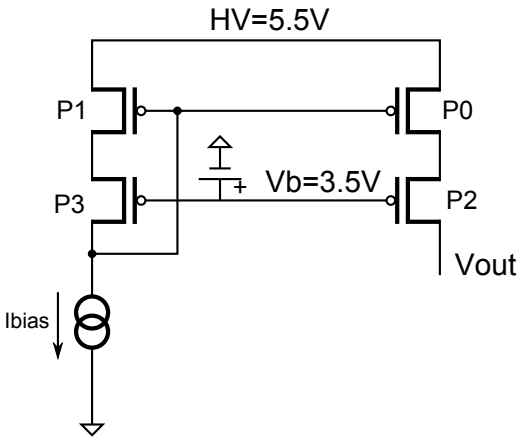


Figure 7.5: Conventional cascode current mirror for high-voltage operation.

7.2.2 Simulations

The DC characteristics I_out-Internal_HV are plotted in figure 7.6. A voltage ramp from 5.5 to 3V is applied to Internal_HV while the output current is plotted for various reference current amplitude.

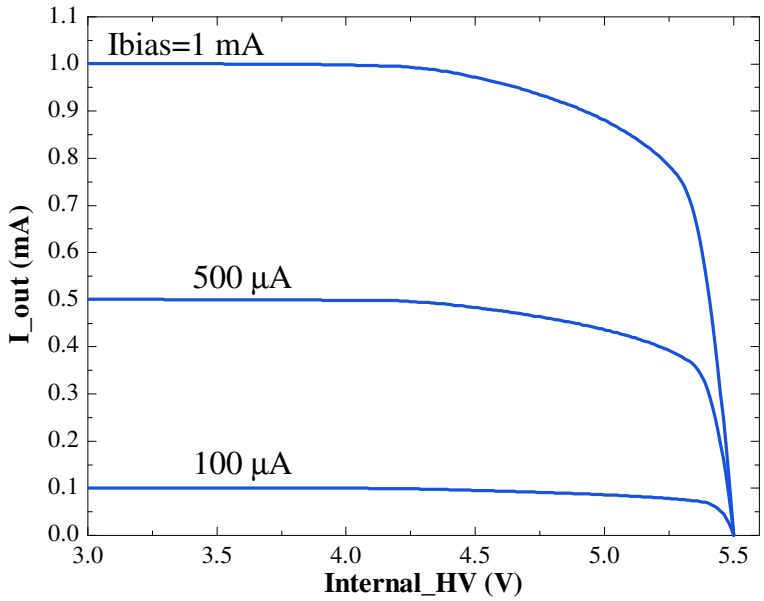


Figure 7.6: I_out-Internal_HV simulations for a reference current of 100μ, 500μ and 1mA.

The maximum programming current can be set from 100μ to 1.3mA as mentioned in table 7.1. The simulation results show that the output current closely tracks the

reference current. However the voltage headroom in the current mirror is about 1.3V for a bias current of 1mA. This limitation is due to the high gate-to-source voltage of P3.
Transient simulations were also performed on the circuit. The antifuse capacitor was replaced by an ideal voltage-controlled resistor. The stray capacitance of the memory array (C_{array}) was extracted and modeled by a capacitor of 1.3pF.
The waveform of the output current is plotted in figure 7.7 for a reference current of 100μ, 200μ and 400μA respectively.

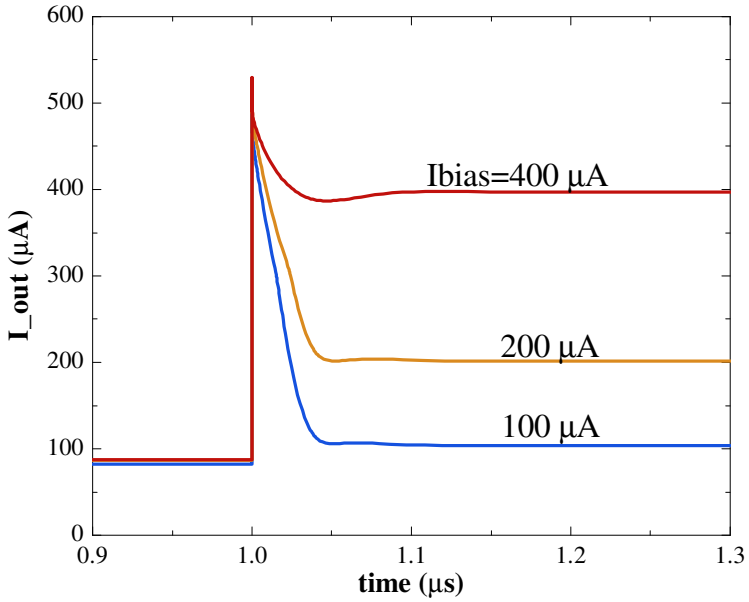


Figure 7.7: Transient simulation of the output current I_out for a reference current of 100μ, 200μ and 400μA. The gate-oxide breakdown is emulated at t = 1μs.

The settling time of the current is significantly reduced compared to the test structure presented in chapter 6. The output current reaches steady state after 50ns whereas 200ns were needed in the single test structure presented in chapter 6.

7.2.3 Implemented solutions

The current source topology presented previously was implemented in the demonstrator. Since the antifuse bitcells are programmed by bit, a single circuit is connected to the memory array. The overhead circuit area is therefore acceptable.
The schematic of the complete circuit implemented in the demonstrator is depicted in figure 7.8.

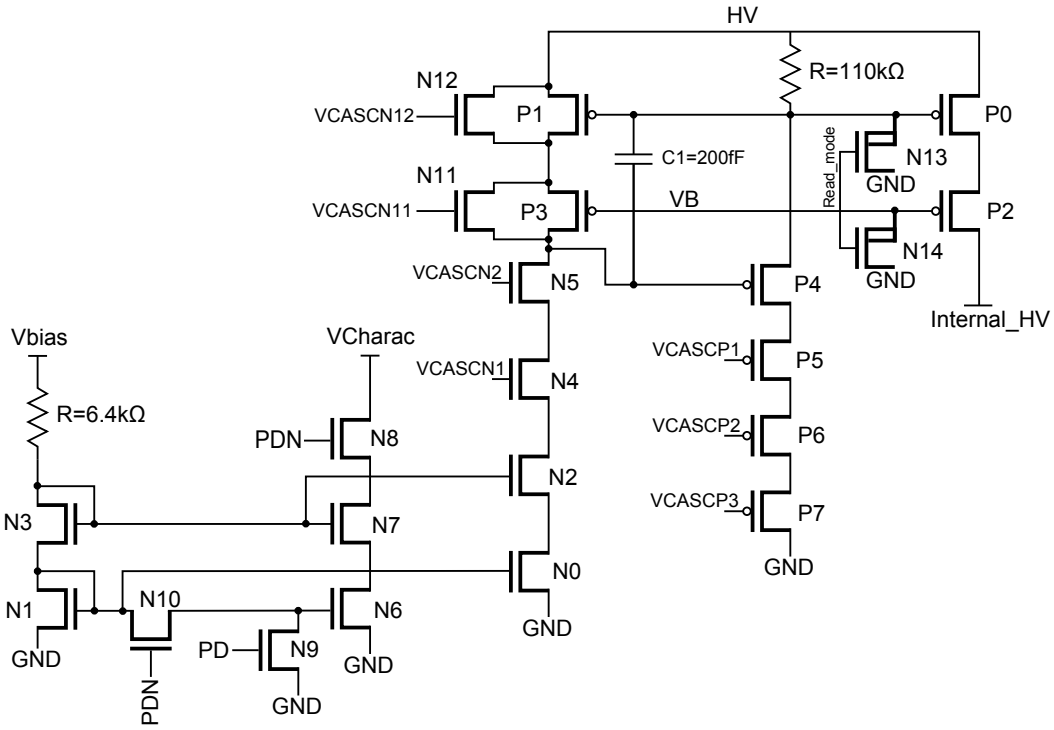


Figure 7.8: Complete schematic of `Programming current limiter` as implemented in the demonstrator.

First an additional branch is added to the current mirror in order to characterize the circuit prior to test. By applying proper signals on `PDN` and `PD`, the reference current range can be characterized using the input `VCharac`.

The transistors `N4` and `N5` are cascoded on top of `N0` and `N2` in order to protect the latter devices from the high programming voltage amplitude. `N11` and `N12` are also protection devices. They avoid the drain-to-source voltage of `P1` and `P3` to reach a higher amplitude than nominal conditions during a ramp on `HV`.

A miller capacitor (C1) was also necessary to insure the stability of the closed-loop system. Finally the drain of `P4` is tied down to the ground by a cascode totem composed of `P5`, `P6` and `P7`.

Two drift transistors `N13` and `N14` are connected to `P0` and `P2` in order to pull down their gates in read mode. As a consequence, the current source is disabled, i.e. the read current is not limited.

7.2.4 Conclusion

The programming current limiter implemented in the antifuse memory demonstrator is an upgraded version of the circuit designed in the single test structure pre-

sented in chapter 6. The cascoded topology enables a significant reduction of the settling time. However, the voltage headroom is higher and may have an impact on the effectiveness of the post-breakdown phase.
The development of a current source compatible with the high programming voltage involves in an antifuse memory and the impedance variation of the breakdown spot is still in a early stage. A test campaign is necessary to assess the performance of the present circuit and to point out possibilities of improvement.

7.3 Programming detection system

Studies presented in chapter 6 provided insights into the contribution of the post-breakdown time on the read current amplitude. However, the latter condition could not be accurately set using the test structure presented in chapter 6.
A configurable programming current sensor and post-breakdown time delay are implemented in the demonstrator in order to detect the breakdown event and to maintain the programming voltage for a given time. The design of this circuit is presented in this section.

7.3.1 Configurable bitline multiplexer

A bitline multiplexer (MUX) is a circuit that routes the bitline of an antifuse bitcell in the memory array to another circuit. The MUX implemented in the demonstrator can be configured either in a standard programming, advanced programming or read mode respectively. A schematic of the bitline multiplexer is depicted in figure 7.9 with the corresponding current path.

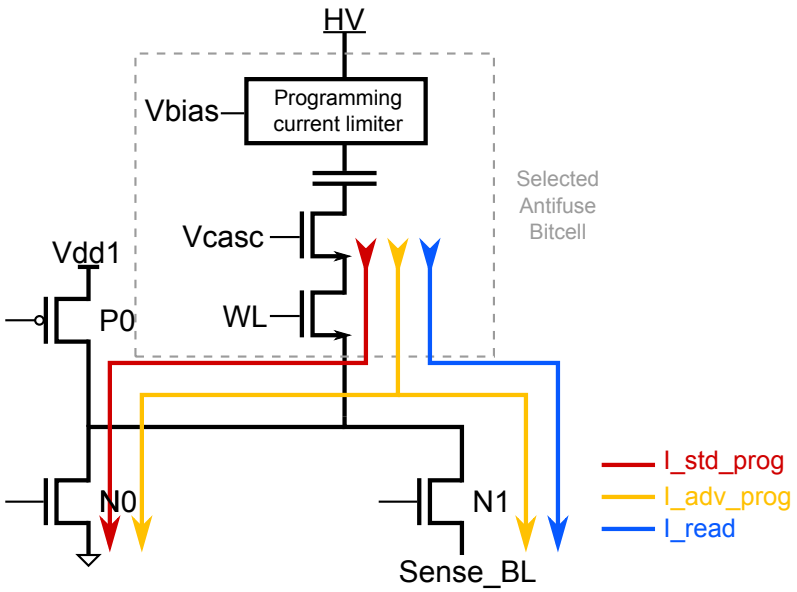


Figure 7.9: Schematic of the configurable bitline multiplexer and the corresponding current path for standard programming, advanced programming and read mode respectively.

In a standard programming mode, the bitline is tied down to the ground by `N0`. The programming current sensor is not used in this mode.
The configuration of the multiplexer in the advanced programming mode involves the transistors `N0` and `N1`. Like in the standard mode, the bitline is tied down to the ground by `N0` while `N1` routes a current-related voltage towards the node `Sense_BL` that is connected to the programming current sensor.
The bitcell is directly connected to the read current sensor using `N1` in read mode. The bitline can also be pulled up to `Vdd1` using `P0`.

7.3.2 Programming current sensor

The programming current sensor is necessary to compare the programming current to a configurable reference in order to detect the gate-oxide breakdown event. A schematic of the circuit is given in figure 7.10.

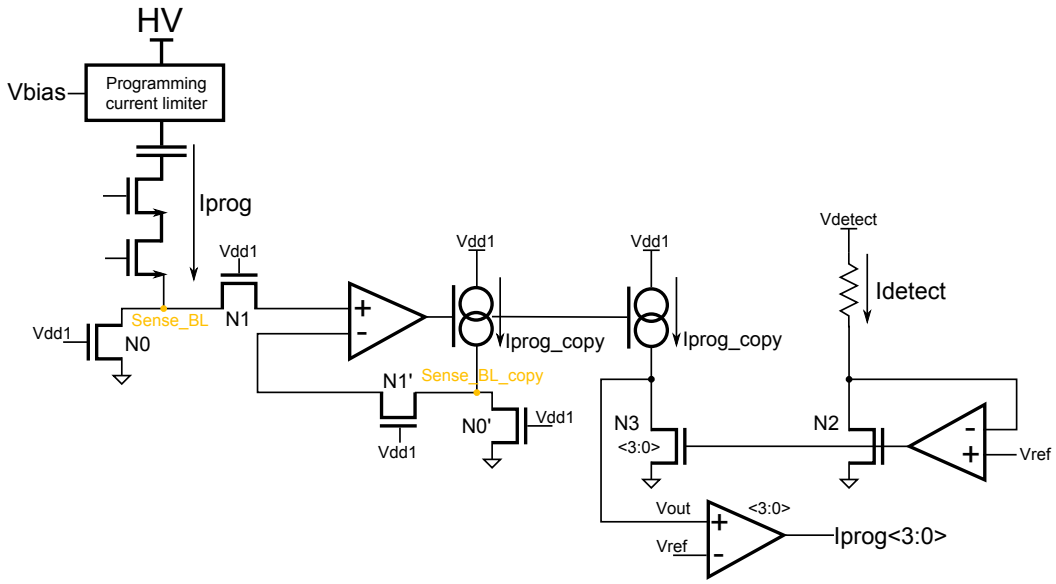


Figure 7.10: Schematic of the programming current sensor connected to the bitline multiplexer and a selected antifuse bitcell.

The selected antifuse bitcell is connected to the programming current sensor using the bitline multiplexer. The bitline is pulled down to the ground using `N0`. The latter transistor is dimensioned such that the drain-to-source voltage amplitude corresponding to the post-breakdown current is around 100mV. Thus, the pass-gate transistor `N1` connects the bitline to the input of an amplifier that features an appropriate input voltage range.

The amplifier drives a voltage-controlled current source connected to the transistors `N0'` and `N1'` that form a structure matched to the bitline MUX (`N0` and `N1`). As a consequence, the programming current flowing through the bitcell is accurately copied in a circuit supplied by Vdd1. Simulations of the bitcell current copy are plotted in figure 7.11.

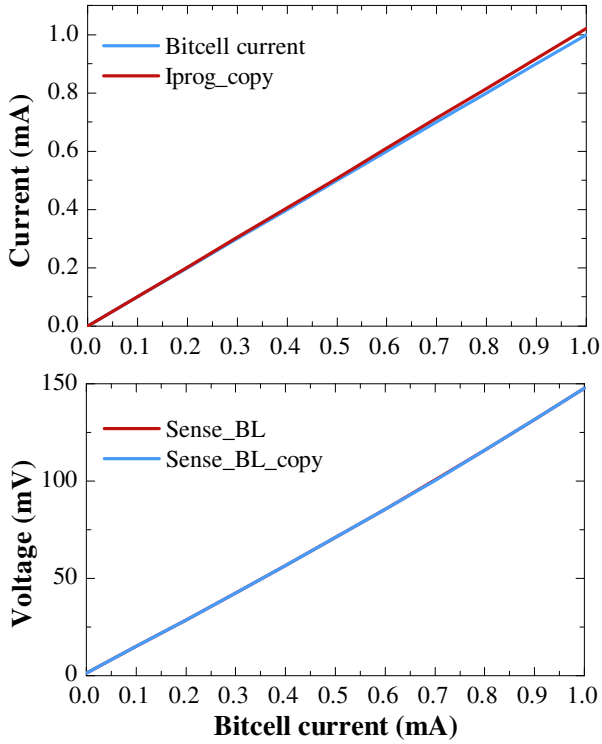


Figure 7.11: Simulation of the bitcell current copy (top) and of the evolution of the voltages Sense_BL and Sense_BL_copy (bottom).

In the simulation testbench, the bitcell is replaced by a current source. A ramp from 0 to 1mA is applied to the `Sense_BL node`. The output current tracks accurately the bitcell current over the whole range. A maximum error of 5% is evaluated in extreme process, voltage and temperature corners.

The evolution of the voltage nodes `Sense_BL` and `Sense_BL_copy` illustrates the low input voltage range of the amplifier. Indeed a maximum voltage of 150mV is simulated for a current of 1mA. Consequently the programming voltage applied to the antifuse bitcell either during the wearout or the post-breakdown phase is not affected by the programming current sensor.

A second current source is driven that is connected to the output of the detection stage. The detection threshold is set by the voltage `Vdetect`. The output voltage of the current mirror formed by `N2` and `N3` is connected to a voltage comparator that detects whether the reference voltage `Iprog_copy` is higher than `Idetect`.

The detection stage features in fact four threshold levels. The information of the threshold level reached by the programming current is carried by the output signal `Iprog<3:0>` and is sent to the post-breakdown time-delay circuit.

A transient simulation of a programming detection is shown in figure 7.12. A current step is applied to `Sense_BL` with a 1ps rise time.

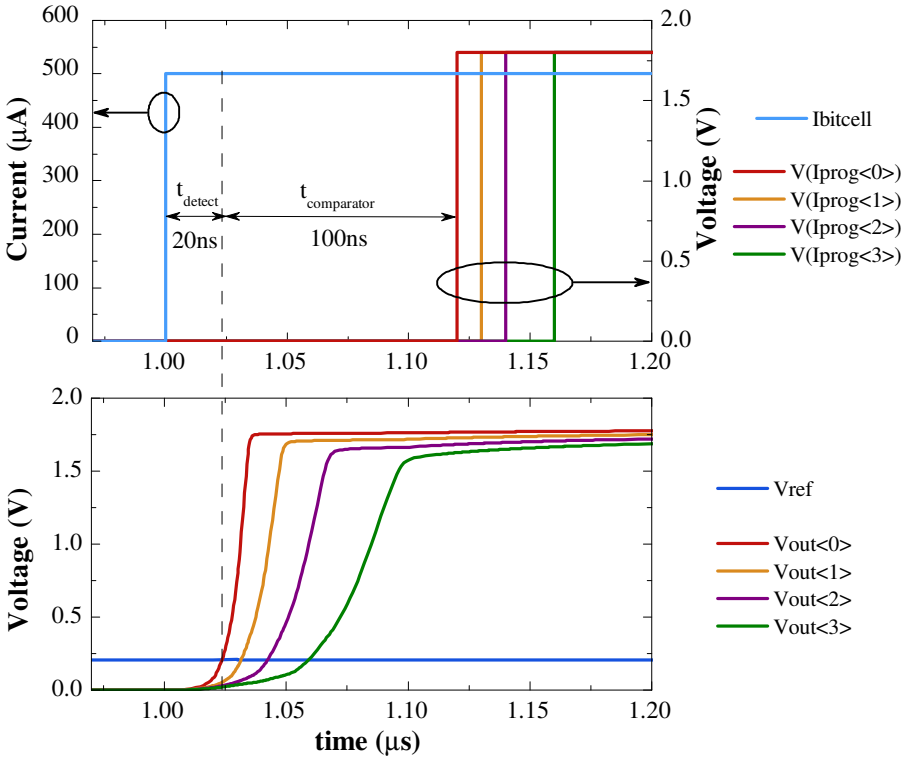


Figure 7.12: Simulation of a programming detection. A current step emulating the gate-oxide breakdown is compared to the output of the voltage comparators (top). The detection occurs when `Vout` becomes higher than `Vref` (bottom).

A current step of 500μA amplitude is applied to `BL` at 1μs. A minimum detection of 20ns is simulated for the lowest detection threshold. Then, the comparator takes 100ns to toggle the output signal `Iprog<0>`. The detection time varies according to the amplitude of the input current and the detection threshold. A detection time range ($t_{detect} + t_{comparator}$) from 66ns to 370ns was simulated over extreme process, voltage, temperature corners and input current variations from 100μA to 800μA.

7.3.3 Post-breakdown time delay

The control of the post-breakdown time is achieved using a frequency divider. Once the breakdown event has been detected, the programming voltage stress is maintained across the antifuse bitcell for a time set using `Tprog<2:0>` and the clock period (`Clk`). The specified range for the post-breakdown time was given in table 7.2. A schematic of the programming detection system is depicted in figure 7.13.

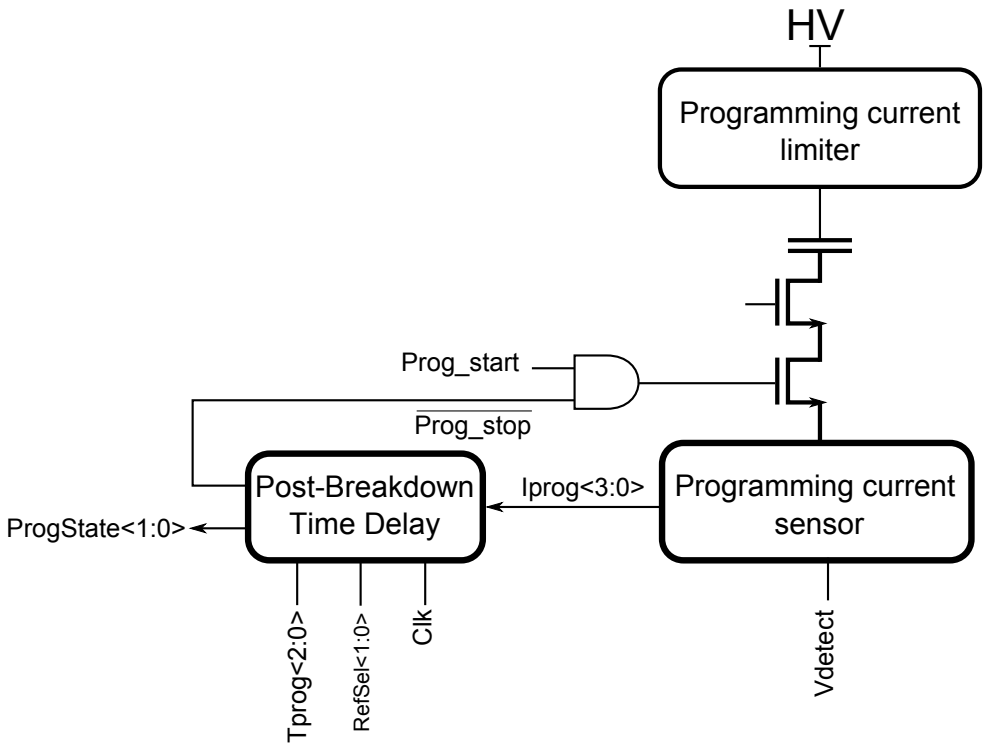


Figure 7.13: Schematic of the programming detection system comprising the current sensor and the post-breakdown time delay.

The post-breakdown time delay is triggered by `Iprog<3:0>` according to `RefSel<1:0>` (see table 7.3). As mentioned previously, the post-breakdown time is set by `Tprog<2:0>` and the clock period (`Clk`).

The output signal `ProgState<1:0>` changes according to the programming sequence as follows:

- `ProgState=00`: Wearout phase.
- `ProgState=01`: Breakdown detected. The post-breakdown time delay is triggered.
- `ProgState=11`: End of the post-breakdown time delay. The programming sequence is over.

The other output signal $\overline{\texttt{Prog\_stop}}$ is connected to a logic gate such that the access transistor is turned off as soon as the post-breakdown time delay has elapsed. A simplified timing diagram was given in figure 7.3.

7.4 Conclusion

The demonstrator presented in this chapter was designed in order to assess the performance of advanced programming modes on a 1-kb antifuse bitcell array. The benefit of a programming current limiter was emphasized in chapter 6. However, the environment of a single bitcell in an augmented test structure is entirely different compared to a device accessed in a dense array. Therefore, the two main goals targeted in the design of the demonstrator are the validation of advanced programming modes on a memory system rather than on an elementary test structure.
Two key features are implemented in the demonstrator. The high programing voltage is applied to the memory array using a current source. The current is thus constant along the post-breakdown phase. A programming detection system was also implemented. It comprises a programming current sensor and post-breakdown time delay. The combinations of these two functions enables the control of the post-breakdown phase in terms of current amplitude and duration.
The configurable memory demonstrator enables the study in a broad range of conditions. The results will complete the previous observations reported in chapter 6.
16 demonstrators have been embedded in a test chip shown in figure 7.14 and taped out on a 32-nm multi-project wafer in December 2010. Hence, 16-kb of antifuse bitcells can be tested in a single die.
Unfortunately the tests have not been performed at the time of this Ph.D. thesis version due to the loaded schedule of the product testing and characterization staff.

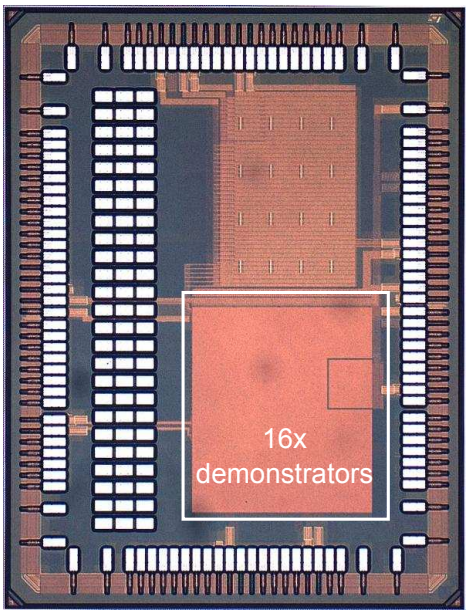


Figure 7.14: Microphotograph of the demonstrator embedded in the test chip taped-out on a 32-nm multi-project wafer.

CHAPTER 8

Conclusion

① The gate-oxide breakdown physics has been studied for decades regarding the reliability of MOS transistors. Although reliability scientists are experts in the physics of dielectric breakdown, the high voltage amplitude used in antifuse memories is not in the range of their study as well as the short time-to-breakdown. The physical mechanisms involved in the gate-oxide breakdown under a high voltage are not known extensively and are not taken into account in the conventional lifetime models.

The objective of this Ph.D. is to investigate the ultrathin dielectric breakdown physics in order to identify and to implement innovative programming methods enabling the reduction of the programming energy. Since a significant circuit area is occupied by the high voltage generator, e.g. a charge-pump circuit, the reduction of the programming voltage and current amplitude has a considerable impact on the latter circuit competitiveness.

❷ The semiconductor memory landscape in the microelectronics industry is presented in chapter 2. The advantage of one time programmable memories are emphasized thanks to the full compatibility with a logic CMOS process. Nowadays, two main technologies are competing: the eFuse and the antifuse. Since the recent CMOS technology nodes feature a thin gate-oxide, the performance of both technologies are comparable. However, the antifuse memory can be used for security applications because the gate-oxide breakdown is difficult to detect using invasive or non-invasive attacks whereas the code stored in an eFuse matrix can be read out using a microscope. Even though the design of antifuse memories is more complex due to the high voltage generator, the security and the robustness are very appealing features for System-on-Chip designers and manufacturers.

③ Chapter 3 focuses on the state-of-the-art of the observations and models of the breakdown of ultrathin gate-oxides (SiO_2). The statistical study of this failure mechanism brought a first model known as the percolation model. The gate-oxide degradation is defined as a stochastic process. As a result, the distribution of the time-to-breakdown for a given stress voltage can be plotted on a Weibull scale.

The voltage-acceleration of the time-to-breakdown is also a major concern. Different modeling approaches are reported and correspond to a gate-oxide thickness range. For current technology nodes, e.g. 45/40-nm, the power-law dependence was proven valid to model the voltage dependence of T_{BD}. Scientists worked on the identification of the physical origin of this dependence. In this respect, a degradation mechanism based on the breakage of Si-H bonds and the release of hydrogen species is reported. Reliability scientists have also reported modeling approaches on high-K dielectric materials as in the 32-nm CMOS process. Due to the composition of the gate stack, the modeling of the breakdown mechanisms is more difficult. Therefore, efforts were focused on fully silicon device in the first place.

A trustful model is necessary to predict the distribution of the time-to-breakdown for a corresponding programming voltage amplitude. Since there are few references regarding this topic, the models identified in reliability must be verified experimentally in the high programming voltage range of antifuse memories. However there is a lack of experimental setup featuring an appropriate bandwidth in order to measure sub-μs time-to-breakdown. Different solutions are discussed in this chapter. A minimum T_{BD} measurement range of 10ns was achieved whereas the minimum T_{BD} range studied in reliability is around 1ms. The validity of the different observations and model can be therefore verified on antifuse bitcells in a high programming voltage range. The characterization methodologies reported in this chapter are also useful to investigate the programming mechanisms on high-K/metal-gate antifuse bitcell.

Preliminary experiments have emphasized 3 different phases occurring in a programming sequence.

1. **Wearout phase:** as soon as a high voltage is applied across the antifuse capacitor, a leakage current flows through the device that damages the dielectric material.

2. **Dielectric breakdown:** a hard breakdown arises.

3. **Post-breakdown:** the high programming voltage is maintained while a high current flows through the breakdown spot.

❹ The wearout phase was investigated in chapter 4. So-called Time Dependent Dielectric Breakdown (TDDB) characterizations were performed in order to model the dispersion of the time-to-breakdown and the voltage dependence.

The measurements performed on various devices emphasized that the leakage current flowing through the gate-oxide before the breakdown event is not negligible and can reach hundreds of μA. As a consequence, a significant voltage drop is lost across the access transistor. Since the time-to-breakdown is accelerated by the effective capacitor voltage, the losses across any series elements must be identified.

There is therefore a need for a wearout current model that enables the calculation of the current amplitude with respect to the programming voltage amplitude and the dimension of the antifuse bitcell. A Fowler-Nordheim model appeared relevant to fit the wearout current in the programming voltage range of antifuse bitcells. As a result, the voltage operating point of the device can be determined as a function of the programming voltage amplitude, the capacitor area and the dimension of the access transistor.

TDDB characterizations were performed in a high voltage range. A first major result is the demonstration of the consistency of the percolation model with the modeling of the distribution of the time-to-breakdown. Constant slopes were identified on Weibull distributions of antifuse bitcells fabricated in a logic 40-nm CMOS process. This property was verified over a wide voltage range. The second step was focused on the identification of the voltage-acceleration law. Power law models identified from measurements performed on a variety of antifuse bitcells with different dimensions were consistent with the acceleration identified in a low voltage range regarding the reliability.

Since T_{BD} is highly dependent on the capacitor voltage, the operating point has a major role in the performance of an antifuse bitcell. This parameter must be optimized by means of maximizing the capacitor voltage during the wearout phase. For this purpose, an analytical model was built. The aim of this model is to calculate the wearout current and the programming voltage amplitude for a targeted T_{BD} as a function of the capacitor area and the dimensions of the access transistor. This methodology yielded guidelines for the optimum design of an antifuse bitcell that reduces the programming voltage amplitude. As a result, a small capacitor area exhibits a lower programming voltage due to an optimized operating point in a high voltage range.

Characterizations were also performed on antifuse bitcells designed and fabricated in a logic 28-nm CMOS process (high-K/metal-gate). The measurements revealed that the insulating stack is less robust and leakier than the silicon dioxide used in 40-nm antifuse bitcells. The number of samples and the lack of knowledge in the

breakdown physics of high-K dielectric materials made difficult the application of the modeling methodology proposed for fully-silicon bitcells in chapter 4. However the importance of the voltage-acceleration and the contribution of the wearout current were emphasized experimentally. Furthermore, the experimental setup developed in this Ph.D. work are relevant for the characterization of this new antifuse bitcell architecture.

The capture of the currents flowing through the bitcell during the breakdown event has highlighted a particular mechanism leading to a current overshoot in the bulk node. The underlying phenomenon was thoroughly investigated in chapter 5. The phenomenon appeared persistent and damaging in terms of power consumption. The characterizations performed on a variety of antifuse bitcells designed an fabricated in different CMOS technology emphasized a dependence of the overshoot amplitude and duration on the bitline current amplitude. Consequently an assumption on the root cause was put on the triggering of a parasitic P-N-P transistor in the antifuse capacitor. The different conditions necessary to operate a bipolar transistor were verified such as the signs of currents, the polarities and the structure. The major concern was the structure of the emitter that is supposed to be a hole reservoir. In the antifuse capacitor, the assumed emitter is the N-type polysilicon gate and the breakdown spot whereas in a conventional P-N-P transistor the emitter is a P+ implant.

TCAD simulations were performed on an antifuse capacitor structure. The study focused on the wearout phase. As a result, it was shown that the high programming voltage leads to a severe band bending and therefore a band-to-band mechanism. Consequently, holes are transported from the gate to the silicon substrate through the gate-oxide.

The latter analysis emphasized that the N-type polysilicon gate can be a source of hole injection and can be considered as the emitter of the parasitic P-N-P bipolar transistor. Furthermore, SPICE simulations performed on a bipolar transistor have confirmed similar caracteristics in terms of current gain with the measurements performed on antifuse bitcell.

⑤ The performance of an antifuse memory is related to the amplitude of the read current, i.e. the characteristic of the breakdown spot. In this respect, chapter 6 deals with the contribution of the post-breakdown phase on the read current amplitude. The post-breakdown current is not limited by the access transistor due to the bulk overshoot. A test structure was designed and implemented with a single bitcell in order to study the contribution of the post-breakdown current amplitude on the read current and the breakdown path characteristic.

Characterizations have shown a dependence of the breakdown path resistance on

the post-breakdown current amplitude. In other words, a qualitative approach emphasized a trade off between the programming and the read current amplitude. The lower the programming current, the lower the read current, however, the programming energy can be significantly reduced.
The contribution of the post-breakdown time could not be studied using the test structure designed for the latter study. An improved structure is therefore needed to further investigate the contribution of the post-breakdown phase on the read current amplitude. Furthermore, quantitative study is study to draw more accurate conclusions.
❻ The final chapter 7 presents an advanced antifuse memory demonstrator of 1-kb designed and fabricated in a logic 32-nm CMOS process. The specifications and the functionalities of this system were defined according to the different studies reported in the previous chapter. The objective is to validate the observations and results on a dense antifuse bitcell array as implemented in a product.
Like in the structure presented in chapter 6, a current limiter is implemented that controls the post-breakdown current. In addition, a programming detection system enables the detection of the breakdown event and the control of the post-breakdown time. The post-breakdown time and current amplitude can be set independently. Results from the future tests will complete the preliminary study reported in chapter 6. The objective is to demonstrate the reduction of the programming energy by limiting the post-breakdown time and/or current amplitude. This demonstrator can be therefore used for electrical characterizations in order to assess the performance of antifuse bitcells according to a variety of programming conditions.
⇨ The work presented here has contributed to the understanding of the gate-oxide breakdown physical phenomenon in a high voltage range that is not covered by conventional reliability studies. The observations and proposed models provide certainties and trends on the appropriate design of antifuse bitcells. Also, the better understanding of the physics and the electrical signatures were essential to propose relevant innovations at system level in order to optimize the programming operation of antifuse bitcell.
⇨ Future work is threefold:

- TCAD simulations with respect to the bulk current overshoot must be confirmed. Then research should focus a technology level solution to annihilate the parasitic P-N-P transistor. The metal gate features also the same problem. This enforces the pertinence of finding a solution.

- The demonstrator will be shortly tested. The work is to assess the operation of the proposed current limiter, the breakdown current detector and the post-breakdown control. Then, it will be possible to conclude on the proposed

assistance to improve the programming of antifuse bitcells for better energy efficiency.

- In a final step, it is interesting to validate the pertinence of the proposed models and assistance circuits in memory arrays designed in more advanced technology nodes.

References

[1] Sidense. (2011, September) Otp applications. Sidense Corporation. [Online]. Available: http://www.sidense.com

[2] Kilopass. (2011, September) Home webpage. [Online]. Available: http://www.kilopass.com/index.php

[3] W. T. Chow, "Storage matrix," US Patent 3 028 659A, 1962.

[4] S. Adee, "Thanks for the memories," *IEEE Spectrum*, vol. 05.09, pp. 44–47, May 2009.

[5] B. Santo, "25 microchips that shook the world," *IEEE Spectrum*, vol. 05.09, pp. 30–39, May 2009.

[6] F. Masuoka, M. Asano, H. Iwahashi, T. Komuro, and S. Tanaka, "A new flash E^2PROM cell using triple polysilicon technology," *Electron Devices Meeting, 1984 International*, vol. 30, pp. 464 – 467, 1984.

[7] B. McClean, *The McClean Report: 2009*, ., Ed. IC insights, 2009.

[8] P. Deroin, "En quête de la mémoire universelle," *L'usine nouvelle*, no. 3193, Mai 2010.

[9] Samsung. (2011, March) Improves the energy performance and increases the savings. [Online]. Available: http://www.samsung.com

[10] Apple. (2011, March) A towering achievement in power. Apple inc. [Online]. Available: http://www.apple.com/macpro/features/processor.html

[11] R. J. Baker, *CMOS Circuit Design, Layout, and Simulation*, 2nd ed. IEEE press, 2005.

[12] S. Jacob, "Intégration, caractérisation et modélisation des mémoires non volatiles à nanocristaux de silicium," Ph.D. dissertation, Université de Provence, April 2008.

[13] F. Piazza, C. Boccaccio, S. Bruyere, R. Cea, B. Clark, N. Degors, C. Collins, A. Gandolfo, A. Gilardini, E. Gomiero, P. Mans, G. Mastracchio, D. Pacelli, N. Planes, J. Simon, M. Weybright, and A. Maurelli, "High performance flash memory for 65 nm embedded automotive application," *Memory Workshop (IMW), 2010 IEEE International*, pp. 1 –3, May 2010.

[14] T. Trexler, "Flash memory complexity," *Instrumentation Measurement Magazine, IEEE*, vol. 8, no. 1, pp. 22 – 26, march 2005.

[15] A. K. Sharma, *Semiconductor Memories*, I. Press, Ed. Wiley-Interscience, Hoboken NJ, 1997.

[16] R. Muralidhar, R. Steimle, M. Sadd, R. Rao, C. Swift, E. Prinz, J. Yater, L. Grieve, K. Harber, B. Hradsky, S. Straub, B. Acred, W. Paulson, W. Chen, L. Parker, S. Anderson, M. Rossow, T. Merchant, M. Paransky, T. Huynh, D. Hadad, K.-M. Chang, and J. White, B.E., "An embedded silicon nanocrystal nonvolatile memory for the 90nm technology node operating at 6V," *Integrated Circuit Design and Technology, 2004. ICICDT '04. International Conference on*, pp. 31 – 35, 2004.

[17] C. Dray and P. Gendrier, "A novel memory array based on an annular single-poly EPROM cell for use in standard CMOS technology," *Memory Technology, Design and Testing, 2002. (MTDT 2002). Proceedings of the 2002 IEEE International Workshop on*, pp. 143 – 148, 2002.

[18] Y. Roizin, E. Pikhay, V. Dayan, and A. Heiman, "High density MTP logic NVM for power management applications," *Memory Workshop, 2009. IMW '09. IEEE International*, pp. 1 –2, May 2009.

[19] A. Atrash, G. Cassuto, W. Chen, V. Dayan, O. Galzur, M. Gutman, A. Heiman, G. Hunsinger, D. Nahmad, A. Parag, E. Pikhay, Y. Roizin, B. Smith, A. Strum, T. Tishbi, and R. Teggatz, "Zero-cost MTP high density NVM modules in a CMOS process flow," *Memory Workshop (IMW), 2010 IEEE International*, pp. 1 –4, May 2010.

[20] R. Bez, S. Bossi, B. Gleixner, F. Pellizzer, A. Pirovano, G. Servalli, and M. Tosi, "Phase change memory development trends," *Memory Workshop (IMW), 2010 IEEE International*, pp. 1 –4, May 2010.

[21] Micron. (2011, March) Flash products webpage. Micron. [Online]. Available: http://www.micron.com/products/

[22] S.-S. Sheu, K.-H. Cheng, M.-F. Chang, P.-C. Chiang, W.-P. Lin, H.-Y. Lee, P.-S. Chen, Y.-S. Chen, T.-Y. Wu, F. Chen, K.-L. Su, M.-J. Kao, and M.-J. Tsai, "Fast-write resistive RAM (RRAM) for embedded applications," *Design Test of Computers, IEEE*, vol. 28, no. 1, pp. 64 –71, 2011.

[23] C. Gopalan, Y. Ma, T. Gallo, J. Wang, E. Runnion, J. Saenz, F. Koushan, and S. Hollmer, "Demonstration of conductive bridging random access memory (CBRAM) in logic CMOS process," *Memory Workshop (IMW), 2010 IEEE International*, pp. 1 –4, May 2010.

[24] N. Raghavan, W. Liu, X. Li, X. Wu, M. Bosman, and K. L. Pey, "Filamentation mechanism of resistive switching in fully silicided high-κ gate stacks," *Electron Device Letters, IEEE*, vol. PP, no. 99, pp. 1 –3, 2011.

[25] N. Xu, B. Gao, L. Liu, B. Sun, X. Liu, R. Han, J. Kang, and B. Yu, "A unified physical model of switching behavior in oxide-based (rram)," *VLSI Technology, 2008 Symposium on*, pp. 100 –101, June 2008.

[26] D. Gogl, C. Arndt, J. Barwin, A. Bette, J. DeBrosse, E. Gow, H. Hoenigschmid, S. Lammers, M. Lamorey, Y. Lu, T. Maffitt, K. Maloney, W. Obermaier, A. Sturm, H. Viehmann, D. Willmott, M. Wood, W. Gallagher, G. Mueller, and A. Sitaram, "A 16-Mb MRAM featuring bootstrapped write drivers," *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 4, pp. 902 – 908, April 2005.

[27] X. Dong, X. Wu, G. Sun, Y. Xie, H. Li, and Y. Chen, "Circuit and microarchitecture evaluation of 3D stacking magnetic RAM (MRAM) as a universal memory replacement," *Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE*, pp. 554 –559, June 2008.

[28] A. Driskill-Smith, S. Watts, D. Apalkov, D. Druist, X. Tang, Z. Diao, X. Luo, A. Ong, V. Nikitin, and E. Chen, "Non-volatile spin-transfer torque RAM (STT-RAM): An analysis of chip data, thermal stability and scalability," *Memory Workshop (IMW), 2010 IEEE International*, pp. 1 –3, May 2010.

[29] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. Kano, "A novel nonvolatile memory with spin torque transfer magnetization switching: spin-ram," *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pp. 459 –462, dec 2005.

[30] K. Lee and S. Kang, "Development of embedded STT-MRAM for mobile System-on-Chips," *Magnetics, IEEE Transactions on*, vol. 47, no. 1, pp. 131 –136, January 2011.

[31] K. Tsuchida, T. Inaba, K. Fujita, Y. Ueda, T. Shimizu, Y. Asao, T. Kajiyama, M. Iwayama, K. Sugiura, S. Ikegawa, T. Kishi, T. Kai, M. Amano, N. Shimomura, H. Yoda, and Y. Watanabe, "A 64Mb MRAM with clamped-reference and adequate-reference schemes," *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, pp. 258 –259, February 2010.

[32] W. Xu and T. Zhang, "Using time-aware memory sensing to address resistance drift issue in multi-level phase change memory," *Quality Electronic Design (ISQED), 2010 11th International Symposium on*, pp. 356 –361, March 2010.

[33] X. Dong and Y. Xie, "AdaMS: Adaptive MLC/SLC phase-change memory design for file storage," *Design Automation Conference (ASP-DAC), 2011 16th Asia and South Pacific*, pp. 31 –36, Jan. 2011.

[34] G. De Sandre, L. Bettini, A. Pirola, L. Marmonier, M. Pasotti, M. Borghi, P. Mattavelli, P. Zuliani, L. Scotti, G. Mastracchio, F. Bedeschi, R. Gastaldi, and R. Bez, "A 4 Mb LV MOS-selected embedded phase change memory in 90 nm standard CMOS technology," *Solid-State Circuits, IEEE Journal of*, vol. 46, no. 1, pp. 52 –63, january 2011.

[35] K. Kim, "Technology challenges for deep-nano semiconductor," *Memory Workshop (IMW), 2010 IEEE International*, pp. 1 –2, May 2010.

[36] J. Lipman, "OTP bit-cell architecture is foundry-friendly," *Chip Design Magazine*, pp. 1–3, September / August 2008.

[37] S. Iyer and W. R. Tonti, "Electrical fuse lets chips heal themselves," *Spectrum, 2004. IEEE*, 2004.

[38] A. Strum, T. Mahlen, and Y. Roizin, "Non-volatile memories in the foundry business," *Memory Workshop (IMW), 2010 IEEE International*, pp. 1 –5, May 2010.

[39] Kilopass. (2011, March) Markets webpage. Kilopass. [Online]. Available: http://www.kilopass.com/markets/

[40] M. Shiozaki, K. Hashimoto, S. Morita, K. Hishioka, and H. Nishimura, "Characterization and optimization of a laser fusing technology for VLSI memories,"

VLSI Technology, 1982. Digest of Technical Papers. Symposium on, pp. 58 – 59, Sept 1982.

[41] K. Arndt, C. Narayan, A. Brintzinger, W. Guthrie, D. Lachtrupp, J. Mauger, D. Glimmer, S. Lawn, B. Dinkel, and A. Mitwalsky, "Reliability of laser activated metal fuses in DRAMs," *Electronics Manufacturing Technology Symposium, 1999. Twenty-Fourth IEEE/CPMT*, pp. 389 –394, 1999.

[42] W. Tonti, "eFuse design and reliability," *Integrated Reliability Workshop Final Report, 2008. IRW 2008. IEEE International*, p. 145, Oct 2008.

[43] Y. Sun, "Laser link cutting for memory chip repair," *Proceedings of the IEEE*, vol. 90, no. 10, pp. 1627 – 1636, Oct. 2002.

[44] M. Alavi, M. Bohr, J. Hicks, M. Denham, A. Cassens, D. Douglas, and M.-C. Tsai, "A PROM element based on salicide agglomeration of poly fuses in a CMOS logic process," *International Electron Devices Meeting, IEDM 1997 Technical Digest.*, pp. 855–858, Dec. 1997.

[45] C. de Graaf, P. H. Woerlee, C. M. Hart, H. Lifka, P. W. H. de Vreede, P. J. M. Janssen, F. J. Sluijs, and G. M. Paulzen, "A novel high-density low-cost diode programmable read only memory," *Electron Devices Meeting, 1996., International*, pp. 189–192, Dec. 1996.

[46] G. Uhlmann, T. Aipperspach, T. Kirihata, K. Chandrasekharan, Y. Z. Li, C. Paone, B. Reed, N. Robson, J. Safran, D. Schmitt, and S. Iyer, "A commercial field-programmable dense eFuse array memory with 99.999% sense yield for 45nm SOI CMOS," *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, pp. 406–407, Feb. 2008.

[47] S. H. Kulkarni, Z. Chen, J. He, L. Jiang, M. B. Pedersen, and K. Zhang, "A 4 Kb metal-fuse OTP-ROM macro featuring a 2 V programmable $1.37\mu m^2$ 1t1r bit cell in 32 nm high-k metal-gate CMOS," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 4, pp. 863–868, 2010.

[48] N. Robson, J. Safran, C. Kothandaraman, A. Cestero, X. Chen, R. Rajeevakumar, A. Leslie, D. Moy, T. Kirihata, and S. Iyer, "Electrically programmable fuse (eFUSE): From memory redundancy to autonomic chips," *Custom Integrated Circuits Conference, CICC 2007, IEEE*, pp. 799–804, Sept. 2007.

[49] C. Kothandaraman, S. Iyer, and S. Iyer, "Electrically programmable fuse (eFUSE) using electromigration in silicides," *Electron Device Letters, IEEE*, vol. 23, no. 9, pp. 523 – 525, Sept. 2002.

[50] T. Ueda, H. Takaoka, M. Hamada, Y. Kobayashi, and A. Ono, "A novel cu electrical fuse structure and blowing scheme utilizing crack-assisted mode for 90-45nm node and beyond," *VLSI Technology, 2006. Digest of Technical Papers. 2006 Symposium on*, pp. 138 –139, 0-0 2006.

[51] S. H. Kulkarni, Z. Chen, J. He, L. Jiang, B. Pedersen, and K. Zhang, "High-density 3-D metal-fuse PROM featuring 1.37μm^2 1t1r bit cell in 32nm high-k metal-gate CMOS technology," *VLSI Circuits Digest of Technical Papers 2009 Symposium on*, pp. 28–29, 2009.

[52] C. Tian, B. Park, C. Kothandaraman, J. Safran, D. Kim, N. Robson, and S. Iyer, "Reliability qualification of CoSi2 electrical fuse for 90nm technology," *Reliability Physics Symposium Proceedings, 2006. 44th Annual., IEEE International*, pp. 392 –397, March 2006.

[53] J. Safran, A. Leslie, G. Fredeman, C. Kothandaraman, A. Cestero, X. Chen, R. Rajeevakumar, D.-K. Kim, Y. Li, D. Moy, N. Robson, T. Kirihata, and I. Subramanian, "A compact eFuse programmable array memory for SOI CMOS," *VLSI Circuits Digest of Technical Papers 2007, Symposiun on*, pp. 72–73, 2007.

[54] R. Degraeve, G. Groeseneken, R. Bellens, M. Depas, and H. E. Maes, "A consistent model for the thickness dependence of intrinsic breakdown in ultra-thin oxides," *Electron Devices Meeting, 1995., International*, pp. 863–866, Dec 1995.

[55] R. Degraeve, G. Groeseneken, R. Bellens, J. L. Ogier, M. Depas, P. J. Roussel, and H. E. Maes, "New insights in the relation between electron trap generation and the statistical properties of oxide breakdown," *IEEE Transactions on Electron Devices*, vol. 45, no. 4, pp. 904–911, Apr. 1998.

[56] P. Candelier, N. Villani, J. P. Schoellkopf, and P. Mortini, "One time programmable drift antifuse cell reliability," *Reliability Physics Symposium, 2000. Proceedings. 38th Annual 2000 IEEE International*, pp. 169–173, Apr. 2000.

[57] H.-K. Cha, I. Yun, J. Kim, B.-C. So, K. Chun, I. Nam, and K. Lee, "A 32-KB standard CMOS antifuse one-time programmable ROM embedded in a 16-bit microcontroller," *Solid-State Circuits, IEEE Journal of*, vol. 41, no. 9, pp. 2115–2124, Sept. 2006.

[58] K. Matsufuji, T. Namekawa, H. Nakano, H. Ito, O. Wada, and N. Otsuka, "A 65nm pure CMOS one-time programmable memory using a two-port anti-

fuse cell implemented in a matrix structure," *Solid-State Circuits Conference, 2007. ASSCC '07. IEEE Asian*, pp. 212–215, Nov. 12–14, 2007.

[59] D. Kaku, T. Namekawa, K. Matsufuji, O. Wada, H. Ito, Y. Sugisawa, S. Shimizu, T. Yamamoto, K. Honda, M. Hamada, and K. Numata, "A field programmable 40-nm pure CMOS embedded memory macro using a PMOS antifuse," *Solid-State Circuits Conference, 2009. A-SSCC 2009. IEEE Asian*, pp. 217–220, Nov. 2009.

[60] J.-P. Son, J. H. Kim, W. S. Ahn, S. U. Han, B.-S. Moon, C. Park, H.-S. Hwang, S.-J. Jang, J. S. Choi, Y.-H. Jun, and S.-W. Kim, "A highly reliable multi-cell antifuse scheme using DRAM cell capacitors," *ESSCIRC, 2010 Proceedings of the*, pp. 482–485, 2010.

[61] J. Peng, G. Rosendale, M. Fliesler, D. Fong, J. Wang, C. Ng, Z. Liu, and H. Luan, "A novel embedded OTP NVM using standard foundry CMOS logic technology," *Non-Volatile Semiconductor Memory Workshop, 2006. IEEE NVSMW 2006. 21st*, pp. 24 –26, Feb 2006.

[62] Kilopass. (2011, March) Security webpage. [Online]. Available: http://www.kilopass.com/technology/cmos-nvm-ip-security/

[63] J. Stathis, "Physical and predictive models of ultrathin oxide reliability in CMOS devices and circuits," *Device and Materials Reliability, IEEE Transactions on*, vol. 1, no. 1, pp. 43 –59, Mar. 2001.

[64] S. M. Sze and K. N. Kwok, *Physics of semiconductor devices*, J. W. . sons, Ed. John Wiley & sons, 2007.

[65] M. Lenzlinger and E. H. Snow, "Fowler-nordheim tunneling into thermally grown SiO_2," *Journal of Applied Physics*, vol. 40, no. 1, pp. 278–283, January 1969.

[66] M. Depas, B. Vermeire, P. Mertens, R. Van Meirhaeghe, and M. Heyns, "Determination of tunneling parameters in ultra-thin oxide layer $Poly - Si/SiO_2/Si$ structures," *Solide-State Electronics*, vol. 38, no. 8, pp. 1465–1471, 1995.

[67] E. Y. Wu, "Ultra-thin oxide realiability for ULSI applications," *Integrated Reliability Workshop Final Report (IRW), 2000 IEEE International. Tutorial Notes*, 2000.

[68] J. Sune, I. Placencia, E. Farres, N. Barniol, and X. Aymerich, "On the breakdown statistics of thin SiO_2 films," *Conduction and Breakdown in Solid Dielectrics, 1989., Proceedings of the 3rd International Conference on*, pp. 364–368, July 1989.

[69] J. Sune, "New physics-based analytic approach to the thin-oxide breakdown statistics," *Electron Device Letters, IEEE*, vol. 22, no. 6, pp. 296 –298, June 2001.

[70] E. Y. Wu and J. Sune, "Power-law voltage acceleration: A key element for ultra-thin gate oxide reliability," *Microelectronics reliability*, vol. 45, pp. 1809–1834, 2005.

[71] P. Nicollian, A. Krishnan, C. Chancellor, and R. Khamankar, "The traps that cause breakdown in deeply scaled SiON dielectrics," *Electron Devices Meeting, 2006. IEDM '06. International*, pp. 1 –4, Dec 2006.

[72] E. Y. Wu, A. Vayshenker, E. Nowak, J. Sune, R. P. Vollertsen, W. Lai, and D. Harmon, "Experimental evidence of T_BD power-law for voltage dependence of oxide breakdown in ultrathin gate oxides," *IEEE Transactions on Electron Devices*, vol. 49, no. 12, pp. 2244–2253, Dec. 2002.

[73] G. Ribes, "Caractérisation et fiabilité des oxydes ultra fins et des diélectriques à forte permittivité issue des technologies cmos 45nm et en deça," Ph.D. dissertation, INP Grenoble, 2005.

[74] J. Sune and E. Y. Wu, "Gate dielectric reliability basic physics and statistics," *Reliability Physics Symposium Tutorial notes, 2009. 47th Annual. 2009 IEEE International*, 2009.

[75] M. Rohner, A. Kerber, and M. Kerber, "Voltage acceleration of T_BD and its correlation to post breakdown conductivity of N- and P-Channel MOSFETs," *Reliability Physics Symposium Proceedings, 2006. 44th Annual., IEEE International*, pp. 76–81, Mar. 2006.

[76] K. F. Schuegraf and C. Hu, "Hole injection SiO2 breakdown model for very low voltage lifetime extrapolation," *IEEE Transactions on Electron Devices*, vol. 41, no. 5, pp. 761–767, May 1994.

[77] J. W. McPherson and H. C. Mogul, "Disturbed bonding states in sio2 thin-films and their impact on time-dependent dielectric breakdown," *Reliability Physics Symposium Proceedings, 1998. 36th Annual. 1998 IEEE International*, pp. pp. 47–56, Mar./Apr. 1998.

[78] J. McPherson, V. Reddy, K. Banerjee, and H. Le, "Comparison of E and 1/E TDDB models for SiO_2 under long-term/low-field test conditions," *Electron Devices Meeting, 1998. IEDM '98 Technical Digest., International*, pp. 171 –174, dec 1998.

[79] J. W. McPherson, "Quantum mechanical treatment of Si-O bond breakage in silica under time dependent dielectric breakdown testing," *Reliability physics symposium, 2007. proceedings. 45th annual. ieee international*, pp. 209–216, Apr. 2007.

[80] D. DiMaria and E. Cartier, "Mechanism for stress induced leakage current in thin silicon dioxide films," *Journal of Applied Physics*, vol. 78, pp. 3883–3894, 1995.

[81] P. Avouris, R. E. Walkup, A. R. Rossi, H. C. Akpati, P. Nordlander, T. C. Shen, G. C. Abeln, and J. W. Lyding, "Breaking individual chemical bonds via STM-induced excitations," *Surface Science*, vol. 363, no. 1-3, pp. 368 – 377, 1996, dynamical Quantum Processes on Solid Surfaces.

[82] G. Ribes, S. Bruyere, M. Denais, F. Monsieur, V. Huard, D. Roy, and G. Ghibaudo, "Multi-vibrational hydrogen release: Physical origin of T_BD, Q_BD power-law voltage dependence of oxide breakdown in ultra-thin gate oxides," *Microelectronics reliability*, vol. 45, no. 12, pp. 1842–54, Dec. 2005.

[83] M. Rafik, "Caractérisation et modélisation des fiabilité des transistors avancés à diélectriques de haute permittivité et à grille métallique," Ph.D. dissertation, INP Grenoble, 2008.

[84] J. Coignus, "Etude le la conduction électrique dasn les diélectriques à forte permittivité utilisés en microélectronique," Ph.D. dissertation, Université de Grenoble, November 2010.

[85] G. Ribes, M. Rafik, D. Roy, and J. Roux, "Reliability issues for nano-scale CMOS dielectrics: - from transistors to product reliability - - from SiON to high-k dielectrics -," *Integrated Circuit Design and Technology and Tutorial, 2008. ICICDT 2008. IEEE International Conference on*, pp. 91 –96, june 2008.

[86] T. Nigam, A. Kerber, and P. Peumans, "Accurate model for time-dependent dielectric breakdown of high-k metal gate stacks," *Reliability Physics Symposium, 2009 IEEE International*, pp. 523 –530, april 2009.

[87] A. Kerber, A. Vayshenker, D. Lipp, T. Nigam, and E. Cartier, "Impact of charge trapping on the voltage acceleration of TDDB in metal gate/high-k N-channel MOSFETs," *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pp. 369 –372, may 2010.

[88] F. Monsieur, "Etude des mécanismes de dégradation lors du claquage des oxydes de grilles ultra minces. application à la fiabilité des technologies cmos sub-0.12μm," Ph.D. dissertation, INP Grenoble, 2002.

[89] F. Monsieur, E. Vincent, G. Pananakais, and G. Ghibaudo, "Wear-out, breakdown occurrence and failure detection in 18-25 åultrathin oxides," *11th Workshop on Dielectrics in Microelectronics (WoDiM 2000)*, vol. 41, no. 7, pp. 1035–9, July 2001.

[90] S.-H. Lee, B.-J. Cho, J.-C. Kim, and S.-H. Choi, "Quasi-breakdown of ultrathin gate oxide under high field stress," *Electron Devices Meeting, 1994. Technical Digest., International*, pp. 605–608, Dec. 1994.

[91] G. Ribes, D. Roy, V. Huard, F. Monsieur, M. Rafik, J. M. Roux, and C. Parthasarathy, "Post breakdown oxide lifetime based on digital circuit failure," *Reliability Physics Symposium, 2008. IRPS 2008. IEEE International*, pp. 215–218, Apr./May 2008.

[92] B. Kaczer, R. Degraeve, M. Rasras, K. Van de Mieroop, P. J. Roussel, and G. Groeseneken, "Impact of MOSFET gate oxide breakdown on digital circuit operation and reliability," *IEEE Transactions on Electron Devices*, vol. 49, no. 3, pp. 500–506, Mar. 2002.

[93] J. Sune, G. Mura, and E. Miranda, "Are soft breakdown and hard breakdown of ultrathin gate oxides actually different failure mechanisms?" *Electron Device Letters, IEEE*, vol. 21, no. 4, pp. 167 –169, apr 2000.

[94] T. Pompl, C. Engel, H. Wurzer, and M. Kerber, "Soft breakdown and hard breakdown in ultra-thin oxides," *Microelectronics reliability*, vol. 41, pp. 543–551, 2001.

[95] B. Linder, S. Lombardo, J. Stathis, A. Vayshenker, and D. Frank, "Voltage dependence of hard breakdown growth and the reliability implication in thin dielectrics," *Electron Device Letters, IEEE*, vol. 23, no. 11, pp. 661 – 663, nov 2002.

[96] M. Alam, B. Weir, and P. Silverman, "A study of soft and hard breakdown - part I: Analysis of statistical percolation conductance," *Electron Devices, IEEE Transactions on*, vol. 49, no. 2, pp. 232 –238, feb 2002.

[97] ——, "A study of soft and hard breakdown - part II: Principles of area, thickness, and voltage scaling," *Electron Devices, IEEE Transactions on*, vol. 49, no. 2, pp. 239 –246, feb 2002.

[98] F. Monsieur, E. Vincent, D. Roy, S. Bruyere, J. Vildeuil, G. Pananakakis, and G. Ghibaudo, "A thorough investigation of progressive breakdown in ultra-thin oxides. physical understanding and application for industrial reliability assessment," *Reliability Physics Symposium Proceedings, 2002. 40th Annual*, pp. 45 – 54, 2002.

[99] Y. Wu, Q. Xiang, D. Bang, G. Lucovsky, and M.-R. Lin, "Time dependent dielectric wearout (TDDW) technique for reliability of ultrathin gate oxides," *IEEE Electron Device Letters*, vol. 20, no. 6, pp. 262–264, June 1999.

[100] J. Wu and E. Rosenbaum, "Gate oxide reliability under ESD-Like pulse stress," *IEEE Transactions on Electron Devices*, vol. 51, no. 9, pp. 1528–1532, Sept. 2004.

[101] A. Kerber, M. Rohner, C. Wallace, L. O'Riain, and M. Kerber, "From wafer-level gate-oxide reliability towards ESD failures in advanced CMOS technologies," *IEEE Transactions on Electron Devices*, vol. 53, no. 4, pp. 917–920, Apr. 2006.

[102] C. Leroux, P. Andreucci, and G. Reimbold, "Analysis of oxide breakdown mechanism occurring during ESD pulses," *Reliability Physics Symposium, 2000. Proceedings. 38th Annual 2000 IEEE International*, pp. 276–282, Apr. 2000.

[103] J. Wu, P. Juliano, and E. Rosenbaum, "Breakdown and latent damage of ultra-thin gate oxides under ESD stress conditions," *Electrical Overstress/Electrostatic Discharge Symposium Proceedings 2000*, pp. 287–295, Sept. 2000.

[104] S. Malobabic, D. F. Ellis, J. A. Salcedo, Y. Zhou, J. J. Hajjar, and J. J. Liou, "Gate oxide evaluation under very fast transmission line pulse (VFTLP) CDM-Type stress," *Devices, Circuits and Systems, 2008. ICCDCS 2008. 7th International Caribbean Conference on*, pp. 1–8, Apr. 2008.

[105] D. F. Ellis, S. Malobabic, J. J. Liou, and J. J. Hajjar, "Prediction of gate dielectric breakdown in the CDM timescale utilizing very fast transmission line pulsing," *Reliability Physics Symposium, 2009 IEEE International*, pp. 585–593, Apr. 2009.

[106] C. D. Young, Y. Zhao, M. Pendley, B. Lee, H., K. Matthews, J. H. Sim, R. Choi, A. Brown, R. W. Murto, and G. Bersuker, "Ultra-short pulse current-voltage characterization of the intrinsic characteristics of high-k devices," *Japanese Journal of Applied Physics*, vol. 44, no. 4B, pp. 2437–2440, April 2005.

[107] E. Miranda, "Method for extracting series resistance in MOS devices using Fowler-Nordheim plot," *Electronics Letters*, vol. 40, pp. 1153–1154, Sept. 2004.

[108] J. W. McPherson, "TDDB physics: Transitioning from silica to High-k gate dielectrics," *Integrated Reliability Workshop Final Report (IRW), 2010 IEEE International*, vol. tutorial note, pp. 168–170, 2010.

[109] B. Razavi, *Design of Analog CMOS Intregrated Circuits*, McGraw-Hill, Ed. McGraw-Hill, 2001.

[110] E. Cartier and A. Kerber, "Stress-induced leakage current and defect generation in nfets with hfo2/tin gate stacks during positive-bias temperature stress," *Reliability Physics Symposium, 2009 IEEE International*, pp. 486 – 492, april 2009.

[111] X. Garros, J. Mitard, C. Leroux, G. Reimbold, and F. Boulanger, "In depth analysis of vt instabilities in hfo2 technologies by charge pumping measurements and electrical modeling," *Reliability physics symposium, 2007. proceedings. 45th annual. ieee international*, pp. 61 –66, april 2007.

[112] J. H. Hur, M.-J. Lee, C. B. Lee, Y.-B. Kim, and C.-J. Kim, "Modeling for bipolar resistive memory switching in transition-metal oxides," *Phys. Rev. B*, vol. 82, no. 15, p. 155321, Oct 2010.

[113] V. L. Lo, K. L. Pey, C. H. Tung, and D. S. Ang, "Effects of nano-scale schottky barrier of conductor-like breakdown path on progressive breakdown in MOSFET," *Reliability Physics Symposium Proceedings, 2006. 44th Annual., IEEE International*, pp. 619–620, Mar. 2006.

[114] J. Martin-Martinez, B. Kaczer, R. Degraeve, P. Roussel, R. Rodriguez, M. Nafria, X. Aymerich, B. Dierickx, and G. Groeseneken, "Circuit design oriented stochastic piecewise modeling of the post-breakdown gate current in MOSFETs: application to ring oscilators," *Device and Materials Reliability, IEEE Transactions on*, vol. PP, no. 99, p. 1, 2011.

[115] A. Ortiz-Conde, E. Miranda, F. J. G. Sanchez, E. Farkas, and S. Malobabic, "Modeling the post-breakdown current in MOS devices on p-silicon substrate," *Devices, Circuits and Systems, Proceedings of the 6th International Caribbean Conference on*, pp. 13–16, Apr. 2006.

APPENDIX A

Synopsis des chapitres

A.1 Introduction générale

La forte demande pour des mémoires non-volatiles embarquées est liée à une émergence de systèmes sur puce de plus en plus complexes. Bien que très denses et très rapides, les mémoires flash sont des solutions coûteuses du fait de leur procédé dédié. La compatibilité des mémoires non-volatiles "programmables une fois" avec des procédés de fabrication CMOS standards leur confère une propriété de bas coût très intéressante pour les fabricants de systèmes sur puce.
Cette thèse traite des mémoires antifusibles qui sont utilisées depuis 10 ans pour de diverses applications telles que : le stockage de code, le stockage de clés d'encryption sécurisées, la traçabilité de circuits ou encore la (re)configuration de systèmes. Malgré une longue période d'utilisation donc une certaine maturité, les possibilités de progrès restent nombreuses. En effet, la programmation embarquée de ces mémoires est rendue possible par un générateur de haute tension de type pompe de charge. L'amplitude de la tension de programmation ainsi que l'énergie nécessaire conduit à un circuit de pompe de charge relativement important et donc d'un coût en surface de silicium non négligeable. La réduction de l'amplitude de la tension de programmation donc de l'énergie mise en jeu permettrait la réduction en surface de la mémoire et, de ce fait, une meilleure compétitivité.

Objectif et contenu de la thèse L'objectif de cette thèse est de proposer des solutions de programmation innovantes afin de réduire l'énergie de programmation: amplitude de la tension, amplitude du courant et la durée. Une connaissance approfondie de la cellule antifusible élémentaire est essentielle pour atteindre cet objectif. C'est pourquoi une partie conséquente de cette thèse traite de l'étude physique du

mécanisme de programmation sous-jacent : le claquage de l'oxyde de grille ultra mince d'une capacité. Une description schématique du courant traversant une capacité pendant une impulsion de haute tension de programmation est donnée figure 1.1 (p. 2). Les différentes étapes du procédé de dégradation du diélectrique sont mises en évidence ainsi que les chapitres relatifs à leur étude.

Le marché lié aux mémoires à semiconducteur est un des plus importants de l'industrie de la microélectronique. Différentes technologies de mémoires sont présentées dans le chapitre 2. Une attention particulière est portée sur les mémoires "programmables une fois".

Un état de l'art sur la physique du claquage d'oxydes ultra mince est détaillé dans le chapitre 3. Les résultats les plus récents concernant la modélisation et l'observation de ce mode de défaillance sont présentés. Le claquage des matériaux diélectriques à haute permittivité est également discuté. Des moyens de caractérisation particuliers sont nécessaires du fait de la haute tension utilisée pour programmer les cellules antifusibles et de ce fait des temps de claquage très courts. Une revue des solutions rapportées dans la littérature est résumée ainsi qu'une évaluation de leur pertinence. Il en résulte de nouveau moyens de caractérisation pour les cellules antifusibles.

Le chapitre 4 présente la caractérisation et la modélisation du temps jusqu'au claquage des cellules antifusibles ainsi que du courant de dégradation. Ces deux paramètres définissent la première phase d'une séquence de programmation. De ce fait, il convient d'évaluer la proportion du temps de programmation occupée par le temps au claquage. Une seconde étape de l'étude consiste à optimiser cette phase en réduisant sa durée et son impact sur la surface occupée par le circuit.

Les caractérisations effectuées sur différentes cellules antifusibles ont révélé un phénomène inattendu déclenché par le claquage de l'oxyde et occasionnant un sur courant substrat très conséquent. Des caractérisations électriques et des simulations ont été effectuées afin de mettre en évidence la cause du mécanisme. Les resultats sont presentés dans le chapitre 5.

La dernière phase de la séquence de programmation se passe après le claquage de l'oxyde et conditionne l'amplitude du courant de lecture. L'étude de celui-ci est rapportée dans le chapitre 6. La conception de structures de test dédiées au contrôle de l'amplitude du courant de programmation est présentée ainsi que les résultats relatifs à la dépendance du courant de lecture avec les conditions post-claquage.

Le chapitre 7 traite de la conception d'un démonstrateur fabriqué avec un procédé CMOS standard 32nm. Ce dernier embarque un plan mémoire de cellules antifusibles d'une capacité de 1kb et des fonctions avancées telles que le contrôle de

l'amplitude du courant de programmation et un système de détection et de temporisation permettant de détecter le claquage et de maitriser le temps pendant lequel la haute tension reste appliquée. Ces fonctionnalités permettent le contrôle indépendant du courant et du temps post-claquage ainsi que l'étude de leur influence sur le courant de lecture sur une grande population statistique et dans un environnement équivalent à un produit industriel.

Les conclusions de cette thèse sont données dans le chapitre 8 ainsi que les perspectives.

A.2 Les mémoires à semiconducteur non-volatiles

Le mémoires à semiconducteur occupent un marché significatif dans le monde de la microélectronique. Les données ainsi que les tendances du graphique de la figure 2.1 (p. 7) montrent que 25% environ du marché total des semiconducteurs est occupé par les memoires. Ce marché solide est donc prévu à un bel avenir car année après année son volume augment ainsi que sa valeur.
Il existe un grand nombre de type de mémoires semiconducteur (voir la figure 2.2, p. 8) et celles-ci adressent des applications différentes. De plus, certaines technologies peuvent être produites comme mémoires de masse. Dans ce cas, la fonction principale est le stockage d'information comme par exemple la mémoire vive d'un ordinateur, une clé USB ou encore un carte mémoire pour appareil nomade. 95 % des revenus présentés en figure 2.1 (p. 7) sont générés par ces mémoires. Les mémoires embarquées sont différentes car elles font partie d'un système sur puce. Un exemple est donné en figure 2.4 (p. 9) où une surface importante d'un microprocesseur multicoeur est occupée par des blocs mémoires. Les pré-requis pour les mémoires de masse ou les mémoires embarqué sont différents. Les performances des technologies Flash, Antifuse et SRAM sont données sur la figure 2.5 (p. 10). Il est montré que chacune possède des avantages et des inconvénients. Par conséquent la mémoire universelle n'existe pas. Le principal inconvénient de la mémoire flash est le coût du procédé et celui de la SRAM, la volatilité des informations stockées. La mémoire antifusible est attractive car elle est non-volatile et compatible avec les procédés standards. Hormis la surface occupée, les coûts des procédés sont faibles. Leur principale limitation réside dans leur programmabilité. Elles ne peuvent ni être effacée ni être reprogrammée. Les mémoires émergentes sont également détaillées et comparées. Il est conclu que bien que prometteuses, aucune d'entre elles n'est en mesure de proposer une solution de mémoire non-volatile à bas coût avec une maturité suffisante.
Les mémoires programmables une fois sont quant à elles présentées dans la section

2.2 (p. 21). Les applications industrielles de ces mémoires sont nombreuses comme listé dans le tableau 2.3 (p. 22). Deux solutions se partagent actuellement le marché : les fusibles et les antifusibles. Un fusible peut etre soit un barreau de poly silicium (voir figure 2.20, p. 26) ou une interconnexion métallique (voir figure 2.21, p. 27). Dans les deux cas, la cellule élémentaire est détruite par un fort courant. Il en résulte par conséquent un circuit ouvert alors qu'un fusible intact présente une très faible résistance. La mémoire fusible présente aujourd'hui un temps de programmation de 1μs par bit pour une surface de cellule de 1.37μm^2. Cependant, la programmation nécessite un courant d'une dizaine de milliampère. Différentes solutions de mémoires fusibles sont résumées dans le tableau 2.5 (p. 28).

Le principe de programmation des mémoires antifusibles repose sur le claquage de l'oxyde d'une capacité. Contrairement au fusible, la cellule est soit très résistive à l'état vierge, soit conductrice à l'état programmé. Le claquage du diélectrique est obtenu en appliquant une haute tension aux bornes de la capacité antifusible. Un schéma de principe illustre le mécanisme sur la figure 2.22 (p. 29). L'amplitude de la tension de programmation est entre 5 à 7 fois supérieure à la valeur nominale de lecture. La cellule élémentaire comprenant une capacité connectée en serie avec un transistor d'accès. Il convient que ce dernier soit capable de supporter la tension de programmation une fois la capacité claquée. La figure 2.23 (p. 30) présente une cellule antifusible en état vierge, programmé et non sélectionné.

Une des clés de la conception d'une cellule antifusible compatible avec un procédé CMOS standard est l'implémentation d'un transistor d'accès adéquate. Différentes architectures sont détaillées dans la section 2.3.2 (p. 29) et leurs performances sont résumées dans le tableau 2.6 (p. 33). Le temps et l'amplitude de la tension de programmation a été sensiblement réduit au cours de ces dix dernières années. La réduction des épaisseurs d'oxyde de grille d'un noeud technologique à un autre permet donc de se rapprocher du temps de programmation des fusibles (10μs) avec un courant de l'ordre de 560μA pour la solution la plus récente.

Les cellules fusible et antifusible présentent des avantages et des inconvénients. Un comparatif est dressé dans le tableau 2.8 (p. 34). Les deux technologies sont compatibles avec un procédé CMOS standard et autorisent la programmation embarquée. L'inconvénient majeur des antifusibles est la haute tension nécessaire à la programmation. Une surface importante est occupée par le circuit de pompe de charge. Deux mémoires fabriquées avec des procédés récents sont comparées en figure 2.28 (p. 35). En comparant les surfaces par cellule, la mémoire fusible présente 9.3μm^2/cellule alors que l'antifusible présente 14μm^2/cellule. Bien que le temps de programmation fût un désavantage certain pour les mémoires antifusibles, l'écart semble se réduire et des temps de programmation inferieur à la dizaine de la

microseconde sont enviseageables.
Malgré des performances légèrement en deçà des mémoires fusibles, la technologie antifusible présente une propriété de sécurité lui permettant d'adresser d'autres domaines d'applications. Contrairement aux fusibles avec lesquels le code stocké peut être lu à l'aide d'un microscope, l'accès à l'information est beaucoup plus difficile avec une mémoire antifusible du fait de la très faible dimension du point de claquage. La figure 2.29 (p. 36) illustre en effet cette propriété.
En conclusion, les mémoires à semiconducteur demeurent un produit stratégique dans l'industrie de la microélectronique. Bien que le marché soit dominé par les technologies DRAM et Flash, des applications spécifiques demandent néanmoins des technologies moins coûteuses pour les mémoires "programmables une fois". Les memoires antifusibles apparaissent comme un candidat intéressant notamment vis-à-vis de la sécurité. Les possibilités d'amélioration n'en demeurent pas moins nécessairess. Une connaissance approfondie des mécanismes de claquage est capitale afin d'identifier des moyens d'améliorations.

A.3 Les mémoires antifusibles et le claquage d'oxyde

L'optimisation des mémoires antifusibles par le biais de la réduction de la surface du circuit et de la consommation n'est possible que si les mécanismes de claquage d'oxyde de grille sont mieux maitrisés. Une illustration simplifiée du courant de programmation d'une cellule antifusible est donnée en figure 3.1 (p. 40). Une approche du phénomène en trois phase est proposée :

- Phase 1 : un courant "d'usure" de porteurs énergétiques traverse le matériau diélectrique.

- Phase 2 : après une durée nommée "temps jusqu'au claquage", le phénomène de claquage se produit. Le courant de programmation augmente depuis le niveau d'usure jusqu'à la saturation du transistor d'accès.

- Phase 3 : le claquage est établi et le courant est stable pendant cette phase post-claquage.

La mesure électrique de ce courant permet la caractérisation du claquage. Ce chapitre se concentre sur la phase 1, c'est-à-dire le courant d'usure et le "temps jusqu'au claquage".
Les différents mécanismes impliqués dans le phénomène du claquage d'oxyde sont illustrés en figure 3.2 (p. 41).

Un des enjeux de ce chapitre est la modélisation comportementale du courant d'usure. Du fait de la pureté des oxydes minces dans les technologies CMOS avancées, des modes de conduction de type tunnel direct ou Fowler-Nordheim peuvent être supposés pour des faibles ou hautes tensions respectivement. Un second enjeu est la modélisation du "temps jusqu'au claquage". Ce domaine est largement couvert dans la littérature concernant la fiabilité des composants. Cependant, les gammes d'amplitude de tension dans lesquelles le claquage est étudié sont sensiblement plus faibles que les amplitudes de tension de programmation rencontrées dans les mémoires antifusibles. C'est la raison pour laquelle la revue d'état de l'art des modèles et des observations du claquage d'oxyde sera confrontée au contexte des mémoires antifusibles.

Le claquage d'oxyde se doit d'être étudié de manière statistique, c'est-à-dire sur une population significative d'échantillons. Ce phénomène étant du point de vue de la fiabilité, une défaillance, il convient de représenter le "temps jusqu'au claquage" par l'intermédiaire d'une distribution de Weibull. Une distribution typique identifiée avec des mesures obtenues de la caractérisation de 72 cellules antifusibles est montrée en figure 3.6 (p. 47). La pertinence de la représentation du "temps jusqu'au claquage" a permis la mise au point d'un modèle incontournable : le modèle de percolation. Celui-ci considère que l'événement du claquage est dû à la génération aléatoire de défauts dans l'oxyde comme illustré en figure 3.8 (p. 49). L'equation (3.13) (p. 49) montre que la pente de la distribution est indépendante de la tension de stress et de la surface du dispositif. De plus le modèle de percolation permet la projection d'une distribution de "temps jusqu'au claquage" d'une surface de capacité à une autre. Cette propriété est illustrée figure 3.10 (p. 50).

La modélisation de la dépendance en tension nécessite une approche plus physique du phénomène en plus du modèle statistique. Une première approche empirique de la dépendance en tension du "temps jusqu'au claquage" a permis de mettre en évidence la pertinence de la loi en puissance aux dépends des lois exponentielles pour les materieau diélectriques ultra minces (voir figure 3.12, p. 53). Cependant, des modèles physiques ont été et sont toujours développés de façon à comprendre l'origine de la loi de puissance et mieux maîtriser la modélisation de la dépendance en tension.

Après plus d'une décennie d'étude sur la fiabilité de composant tout silicium (oxyde SiO_2), un changement drastique est intervenu dans les technologies avancées 32 et 28nm. En effet le diélectrique de dioxyde de silicium est remplacé par un matériau à haute permittivité comme illustré sur la figure 3.18 (p. 60). Bien que les méthodes de caractérisations restent valides, le changement de composition des composants induits de nouveaux phénomènes, difficiles à modéliser. L'approche statistique est

toujours valable ainsi que l'accélération en tension du stress. La difficulté demeure pour établir un modèle de fiabilité couvrant une large gamme dimensionnelle de composants. La conception et le fonctionnement de cellule antifusible disposant d'un matériau diélectrique haute permittivité ne parait néanmoins pas compromise.

Comme expliqué précédemment, les modèles statistiques ou physiques développés pour la fiabilité n'ont pas été validés dans le domaine de la haute tension et des "temps jusqu'au claquage" très court (1μs ou inferieur). Des moyens de caractérisation particuliers sont nécessaires pour mesurer des "temps jusqu'au claquage" de l'ordre de cette durée et donc pour pouvoir verifier la validité des modèles. C'est la raison pour laquelle un état de l'art des techniques de mesure de courant est documenté. Il convient d'identifier des protocoles expérimentaux prometteurs dans le but de mener des campagnes de caractérisation.

L'observation de l'évolution du courant de grille d'une capacité a permis de constater différents modes de défaillance. En effet un claquage dur est représenté par un saut de courant très abrupte alors que le claquage progressif est reflété par un courant bruité et une évolution plus douce. Le premier phénomène est plutôt observé sur des oxydes épais et des hautes tensions alors que le second est rencontré dans les oxydes fins et des tensions plus faibles. Ces deux modes sont illustrés sur la figure 3.22 (p. 64). Les techniques de caractérisation misent en place pour les cellules antifusibles permettront donc de statuer sur les modes de claquage. Différentes techniques de caractérisation sont présentées dans ce chapitre et deux sont particulièrement prometteuses. La mesure sur tranche peut être réalisée à l'aide d'une résistance série et d'un oscilloscope. L'utilisation des entrées 50-Ω permet la réduction de la longueur des câbles tout en gardant un bon compromis bande-passante/précision. Un "temps jusqu'au claquage" minimum de 1μs peut être mesuré avec cette technique. Cette méthode permet aussi la mesure d'amplitude des courants d'usure et post-claquage.

La bande passante peut encore être améliorée en utilisant une cellule antifusible montée sur un plot RF. Le courant de programmation est dans ce cas mesuré à l'aide d'un *Bias-Tee* sur un banc de test RF. Cette amélioration permet la mesure de "temps jusqu'au claquage" inferieur à 10ns.

Ce chapitre a permis de mettre en évidence les travaux nécessaires à la compréhension et à la modélisation du claquage d'oxyde de grille. Les moyens expérimentaux sont maintenant au point et vont permettre de verifier la validité des modèles de fiabilité ainsi que les modes de claquages dans une gamme de haute tension.

A.4 Modélisation TDDB pour la conception de cellules antifusibles

Ce chapitre se concentre sur l'étude de la première phase du claquage d'oxyde. C'est-à-dire de l'application de la tension de programmation jusqu'au claquage. Des caractérisations *Time-Dependent Dielectric Breakdown* (TDDB) sont réalisées sur des cellules antifusibles fabriquées dans une technologie CMOS standard 45nm et 40nm. Des résultats sur des cellules fabriquées avec un procédé CMOS avancé 32nm sont également présentés en fin de chapitre.

Un exemple de formes d'ondes de courant est donné en figure 4.2 (p. 81) pour illustrer la méthodologie de caractérisation. Il est montré que l'amplitude du courant d'usure est importante sous haute tension. Comme la capacité antifusible est connectée en série avec un transistor d'accès, la chute de tension aux bornes de celui-ci doit être prise en compte afin d'estimer la tension effective aux bornes de la capacité. Le claquage étant accéléré par la tension, un "temps jusqu'au claquage" court sera garanti en minimisant la chute de tension aux bornes du transistor d'accès. Deux exemples de points de fonctionnement sont donnés en figure 4.5 (p. 84).

De part son amplitude, le courant d'usure joue un rôle important dans cette première phase. Un modèle comportemental est par conséquent utile afin de déterminer précisément la tension effective appliquée aux bornes de la capacité avant l'événement du claquage. Du fait de l'importante amplitude de la tension de programmation, un modèle de conduction de type Fowler-Nordheim s'est avéré pertinent. Les paramètres du modèle ont été identifiés d'après des caractérisations TDDB préalables comme illustré sur la figure 4.8 (p. 87). Puis, la projection d'une surface de capacité à une autre a été vérifiée (voir figure 4.9, p. 89).

La caractérisation et la modélisation du "temps jusqu'au claquage" a aussi été traitée dans le chapitre. Des mesures ont été effectuées sous différentes tensions de programmation afin de déterminer les distributions de Weibull correspondante. Une gamme de tension de 3.4 à 7V a été couverte pour des temps médians correspondant d'environ 1000s à 100ns. Les distributions sont reproduites sur les figures 4.11 et 4.13 (p. 92 et 95). Ces premiers résultats montrent que la pente de Weibull de toutes les distributions sont indépendantes de la tension de programmation et de la surface de la capacité. Cette propriété du modèle de percolation est donc validée pour des cellules antifusibles programmées sous des hautes tensions.

Grâce au modéle de courant d'usure identifié précédemment, les tensions effectives appliqués aux bornes des capacités antifusibles avant le claquage peuvent être déterminées. Les temps médians des distributions peuvent donc être tracés en fonction de cette tension afin d'identifier une loi d'accélération du "temps jusqu'au

claquage" en fonction de l'amplitude de la tension appliquée aux bornes de la capacité. Les résultats sont montrés sur les figures 4.12 et 4.14 (p. 93 et 96). Des lois en puissance présentant un facteur d'accélération de -43 ± 3 ont été identifiées sur différents dispositifs. Ce résultat est en adéquation avec les valeurs rapportées dans la littérature. La loi en puissance semble donc adapter pour décrire le "temps jusqu'au claquage" des cellules antifusibles dans un domaine de haute tension. De plus un modèle peut être projeté d'une surface de capacité à une autre grâce au modèle de percolation.

La maîtrise de la modélisation de l'amplitude du courant d'usure ainsi que de la dépendance du "temps jusqu'au claquage" en fonction de la tension appliquée aux bornes de la capacité antifusible permet d'appréhender les effets des dimensions de la capacité et du transistor d'accès sur le "temps jusqu'au claquage". Un circuit équivalent à la cellule antifusible est proposé en figure 4.15 (p. 98). Une équation de la tension aux bornes de la capacité peut être construite en fonction de la tension de programmation, de l'amplitude du courant d'usure de la résistance du transistor d'accès quand celui-ci est en régime de fonctionnement ohmique. Cette approche de modélisation est illustrée par le diagramme de la figure 4.16 (p. 99). Finalement, le "temps jusqu'au claquage" et le courant d'usure peuvent être calculés en fonction de l'amplitude de la tension de programmation. Les différents paramètres des deux modèles doivent néanmoins être identifiés au préalable par des caractérisations.

Ce modèle est appliqué à des cellules antifusibles ayant des surfaces de capacités différentes mais un transistor d'accès similaire. La méthode de modélisation peut donc être vérifiée. Les résultats sont montrés sur les figures 4.18, 4.19 et 4.20 ((p. 101, 102 et 103)). Dans le domaine de la haute tension, le "temps jusqu'au claquage" le plus faible est obtenu pour la cellule antifusible ayant la plus petite surface. Ce résultat est intéressant car selon le modèle de percolation, un "temps jusqu'au claquage" long est obtenu pour une petite capacité (à tension équivalente). C'est d'ailleurs le cas en faible tension. Ce changement de tendance est en fait expliqué en étudiant le point de fonctionnement des différentes cellules. Un exemple pratique est illustré sur la figure 4.21 (p. 104). Une cellule ayant une faible surface de capacité présente une amplitude de courant d'usure plus faible et par conséquent une tension aux bornes de la capacité antifusible plus importante. La dépendance du "temps jusqu'au claquage" avec la tension est régi par une loi de puissance d'un facteur très important alors qu'une relation linéaire relie la surface au "temps jusqu'au claquage". C'est la raison pour laquelle la tension effective aux bornes de la capacité est le facteur dominant.

La validation de cette méthode de modélisation prenant en compte les dimensions de la cellule antifusible permet de mettre au point l'algorithme d'optimisation de

la figure 4.24 (p. 108). Ce dernier permet de calculer l'amplitude de la tension de programmation en fonction d'un "temps jusqu'au claquage" donné et des dimensions de la capacité et du transistor d'accès. Les résultats sont montrés sur les figures 4.25 et 4.26 (p. 109 et 110). Un dimensionnement de cellule approprié permet de réduire considérablement la tension de programmation. Cette tendance est d'autant plus vraie pour des temps courts jusqu'au claquage.

Le travail de caractérisation et de modélisation a permis une meilleure compréhension de tenants et aboutissants de cette première phase d'une séquence typique de programmation. La réduction du temps total de programmation passe en effet par un "temps jusqu'au claquage" court. La tendance montrant les bénéfices d'une capacité de faible surface est intéressante d'autant que celle-ci est valable si des "temps jusqu'au claquage" courts sont recherchés.

Des caractérisations de cellules antifusibles fabriquées sur une technologie CMOS avancée 28nm ont également été réalisées. Il a été observé que le matériau diélectrique est plus fragile que le SiO_2 conventionnel. Le courant de fuite est par contre plus important comme le montre la figure 4.29 (p. 115). Les performances de "temps jusqu'au claquage" sont néanmoins très prometteuses (voir la figure 4.30, p. 116). Bien que la modélisation TDDB de ce type de cellule n'a pas été abordé, les méthodes de caractérisation développées pour les noeuds technologique 45 et 40nm peuvent être utilisées sur ces nouveaux dispositifs.

En conclusion, il est intéressant de noter que le "temps jusqu'au claquage" est très court vis-à-vis du temps de programmation total. Bien que la compréhension et la modélisation de cette phase soit nécessaire, la plus grande partie de l'énergie de programmation est consommée pendant la phase post-claquage. Cette phase est abordée au chapitre 6.

A.5 Effet secondaire : sur-courant de substrat

Les caractérisations TDDB présentées dans le chapitre précédent ont permis de mettre en évidence un sur-courant de substrat déclenché lors du le claquage de l'oxyde. Le courant de programmation n'est plus limité par le transistor d'accès. Il convient donc de comprendre le mécanisme occasionnant ce sur-courant afin de limiter son effet sur la puissance consommée. Le schéma de principe du banc de test ainsi que des formes d'ondes typiques sont montrées sur les figures 5.1 et 5.2 (p. 121 et 122).

Le phénomène a été étudié par l'intermédiaire de caractérisations électriques de cellules antifusibles d'architecture *drift*. Un des premiers résultats a montré que la cause du mécanisme est liée à la capacité antifusible. Le courant de porteur chaud

du transistor d'accès ne présente pas une amplitude suffisante.

L'impact des conditions de programmation a ensuite été étudié. Des caractérisations électriques ont été effectuées sous différentes tensions de programmation et différentes tensions de *Word Line* permettant de changer l'amplitude du courant de saturation du transistor d'accès. Des résultats sont montrés sur les figures 5.11 et 5.13 (p. 133 et 135). Un lien direct existe entre l'amplitude du courant de saturation du transistor d'accès et les caractéristiques du sur-courant de substrat. Le phénomène est en effet atténué en terme d'amplitude et de durée si le courant circulant dans le canal du transistor d'accès est diminué. Le mécanisme est donc également atténué dans le cas où la tension de programmation n'est pas suffisante pour permettre la saturation du transistor d'accès après le claquage de l'oxyde. Des expériences additionnelles sous différentes conditions de température ou avec différentes tensions de substrat ont été menées sans pouvoir directement modifier le courant (les effets du courant de saturation du transistor d'accès étant mis à part). Le tableau 5.2 récapitule les résultats clés des caractérisations électriques.

Le phénomène occasionnant le sur-courant de subsrtat est persistant et problématique. Il est donc important de mettre en évidence la cause de ce mécanisme afin de l'éliminer. Le trajet des électrons et des trous dans la cellule antifusible a été analysé pendant la phase d'usure et post-claquage.

Bien que le courant substrat soit relativement faible pendant la phase d'usure, il est tout de même supérieur au courant de porteurs chauds du transistor d'accès. Il est donc possible que ce courant soit précurseur du phénomène amplifié par le claquage. La vue en coupe de la cellule antifusible en figure 5.14 (p. 138) montre que des électrons transitent par le canal du transistor avant de s'accumuler sous la grille. Le courant de substrat est par contre dû à des trous. Des simulations TCAD ont été effectuées afin d'étudier les courbures de bandes dans la cellule antifusible et identifier le ou les mécanismes responsables du courant de substrat. Le diagramme de bandes sous différentes tensions de programmation est montré à la figure 5.15 (p. 139). Les bandes d'énergie sont fortement courbées du fait de la haute tension appliquée sur la cellule. Dans ces conditions, une injection de trous de la grille vers le substrat est possible par un mécanisme de conduction bande à bande. Les simulations TCAD de courant de grille et de substrat tracées à la figure 5.16 (p. 140) montrent des ordres de grandeur conformes aux mesures électriques. Un paramétrage plus précis des modèles serait nécessaire pour obtenir des simulations plus précises. La mise en évidence de la provenance des trous et du mécanisme de conduction sont néanmoins des résultats précieux.

Le trajet des électrons et des trous a aussi été étudié pendant la phase post-claquage comme le montre la figure 5.17 (p. 141). Cependant des simulations ne peuvent

être réalisées du fait de la difficulté de modélisation de l'oxyde claqué. Malgré les différences d'amplitudes de courant, le trajet des porteurs est similaire. L'avalanche de la jonction Nwell/P-substrat est facilement écartée par une simulation SPICE de ce type de jonction. Comme montré sur la figure 5.18 (p. 142), l'avalanche est déclenchée pour une tension supérieure à 14V.

Un autre mécanisme possible serait le déclenchement d'un transistor bipolaire parasite situé dans la capacité antifusible. Le collecteur serait le substrat de type P, la base le caisson Nwell et l'émetteur la grille. le point de claquage est illustré sur la figure 5.19 (p. 143). La structure, le sens des courants ainsi que les polarités concordent avec ceux d'un transistor bipolaire de type P-N-P. De ce fait, le courant substrat est assimilé au courant collecteur et le courant du canal du transistor d'accès au courant de base. Le courant de grille de la capacité étant le courant d'émetteur, la somme des courant collecteur base est bien retrouvée.

La structure de l'émetteur mise à part, les transistors bipolaires utilisés dans les technologies CMOS 45 et 40nm ont des dimensions similaires. Des simulations SPICE permettent de retrouver les grandeurs électriques mesurées électriquement sur les cellules antifusibles. Les variations de gain en fonction de l'amplitude du courant de base sont en effet similaires. Les mesures électriques et les simulations sont montrées sur les figures 5.24, 5.25 et 5.26 (p. 148, 149 et 150).

La structure de l'émetteur dans la capacité antifusible peut prêter à question. La grille de polysilicium étant de type N, il est nécessaire de prouver que celle-ci est capable d'émettre les trous nécessaires au déclenchement du transistor bipolaire. Les simulations TCAD présentées précédemment permettent de justifier cette hypothèse.

L'hypothèse du déclenchement d'un transistor bipolaire PNP parasite situé dans la capacité antifusible parait être la plus probable pour justifier le mécanisme conduisant à une sur-courant de substrat lors de la programmation. L'atténuation du courant parfois brutale reste néanmoins à expliquer.

Ce mécanisme est persistant et son élimination ne parait pas triviale. Il est nécessaire de pouvoir contrôler le courant passant dans le point de claquage afin d'étudier la contribution de la phase post-claquage sur le courant de lecture. C'est l'objet du chapitre suivant.

A.6 La phase post-claquage et le courant de lecture

L'étape finale de la programmation est la phase post-claquage. Celle-ci est définie sur la figure 6.1 (p. 154) par un temps mesuré entre l'événement du claquage et la fin de l'impulsion de programmation. L'enjeu de ce chapitre est de comprendre la

contribution de cette phase sur l'amplitude résultante du courant de lecture.

La lecture d'une cellule antifusible est réalisée en appliquant une tension faible (n'altérant pas le matériau diélectrique) sur la grille de la capacité et en mesurant le courant de fuite. Ainsi un état vierge ou programmé peut être identifié comme le montre la figure 6.2 (p. 155). En plus de l'amplitude du courant, la caractéristique courant-tension d'une cellule antifusible programmée permet d'identifier les composantes possibles du point de claquage. Les mesures électriques montrées en figure 6.3 (p. 156) révèlent que le point de claquage est en parti résistif. Dès lors un circuit équivalent est proposé à la figure 6.5 (p. 158).

Pour étudier la contribution du courant post-claquage sur le point de claquage et donc sur le courant de lecture, il convient de maîtriser l'amplitude du courant une fois que le claquage est apparu. Comme expliqué dans le chapitre précédent, le transistor d'accès ne limite pas le courant post-claquage du fait du sur-courant de substrat comme illustré figure 6.6 (p. 160). Une source de courant est donc implémentée avec une cellule élémentaire antifusible. Le schéma de la structure de test est donné à la figure 6.7 (p. 161).

Cette source de courant permet de contrôler le courant maximum passant dans le point de claquage. Des mesures statique et dynamique sont montrées figure 6.8 et 6.9 (p. 162). Des expériences de programmation sous différentes amplitudes de courant post-claquage. Des distributions de courant de lecture sont tracées à la figure 6.10 (p. 164). La valeur moyenne des distributions est en effet influencée par l'amplitude du courant post-claquage. En effet, un courant de lecture plus faible est obtenu pour un courant post-claquage plus faible. Les cellules programmées ont ensuite été caractérisées. Les caractéristiques courant-tension sont tracées à la figure 6.11 (p. 166). Les valeurs de résistance ont également été identifiées. Celles-ci sont impactées par les conditions sur le courant post-claquage. Des distributions sont montrées figure 6.13 (p. 168). Le rôle de la partie résistive du point de claquage est donc mis en évidence par ces caractérisations. Si cette méthodologie de modélisation présente des qualités descriptives, elle ne peut pas être utilisée pour le moment pour aider à la conception.

Cette structure de test a également été conçue avec une cellule antifusible de type "cascode" afin de tester des programmations avec des amplitudes de courant post-claquage plus faibles et une plus grande amplitude de tension. Les distributions sont montrées figure 6.14 (p. 169). Des amplitudes de courant de lecture relativement élevées ont été obtenues pour des niveaux de courant de programmation très faibles. Il a été montré que le temps d'établissement de la source de courant est d'environ 200ns (voir les formes d'ondes sur la figure 6.15, p. 170). Pendant cette période transitoire, le courant post-claquage n'est pas encore limité et les porteurs sont

tres énergétiques. L'evolution des courant et des potentiels dans la structure de test est illustré figure 6.16. Il est donc probable que le courant de lecture soit fixé pendant cette phase. Une fois le courant établi, la chute de potentiel aux bornes de la source de courant est telle que les porteurs ne possèdent plus l'énergie suffisante pour dégrader le point de claquage.
Des résultats qualitatifs quant à la dépendance du courant de lecture avec les conditions post-claquage ont été mis en évidence dans ce chapitre. Les caractérisations ont pu être réalisées grâce à une structure de test dédiée au contrôle du courant post-claquage. Des axes d'amelioration ont été identifiés pour reduire le temps d'etablissement. Ce système sera donc porté dans un système de mémoire antifusible afin de consolider les précédents résultats et proposer une étude quantitative. La conception de ce système est presentée dans le chapitre suivant.

A.7 Démonstrateur de mémoire antifusible en CMOS 32nm

Le dernier chapitre de cette thèse présente la conception d'un démonstrateur de mémoire antifusible avec un procédé CMOS standard 32nm. L'objectif est d'utiliser les connaissances sur les mécanismes de claquage rassemblées dans les chapitres précédents afin de proposer une validation des fonctions de programmation avancées au niveau système.
Le schéma de principe du démonstrateur est proposé à la figure 7.1 (p. 177). Il se compose d'un plan mémoire de 1kb accédé aléatoirement, d'une source de courant permettant de limiter le courant post-claquage, d'un système de détection de programmation et d'une base de temps permettant de contrôler le temps post-claquage et d'un système de lecture. Le tout est contrôlé par un bloc logique.
Deux modes de programmation sont possibles. Le mode standard est illustré figure 7.2 (p. 178). La programmation est contrôlée par une impulsion calibrée. Le système de détection n'est pas utilisé dans ce mode. Le mode de programmation avancé est illustré à la figure 7.3 (p. 179). Celui-ci permet de détecter le claquage et de contrôler le temps post-claquage. Le courant est également limité. De cette façon la contribution de la phase post-claquage sur le courant de lecture peut être étudiée en termes de temps et d'amplitude de courant. De plus, l'étude est réalisée sur des cellules dans un environnement mémoire et avec une plus grande population statistique. L'étude quantitative sera donc possible et constitue un objectif majeur. Les spécifications du démonstrateur sont listées dans le tableau 7.1 (p. 180).
La conception de la source de courant est détaillée dans la section 7.2 (p. 181). Il s'agit d'une topologie améliorée par rapport à la structure de test présentée dans

le chapitre précédent. Le schéma est décris dans les figures 7.4 et 7.8 (p. 182 et 185). Les simulations montrées aux figures 7.6 et 7.7 (p. 183 et 184) illustrent les performances de cette source de courant.
Le système de détection de programmation est également présenté. Le schéma est décrit à la figure 7.10 (p. 188). Ce circuit permet de copier le courant circulant dans la cellule antifusible pendant la programmation et de le comparer à un seuil fixé au préalable. L'objectif est de détecter le front de courant signalant l'événement du claquage. Des simulations montrant les performances du système sont disponibles sur les figures 7.11 et 7.12 (p. 189 et 190). Une fois le claquage détecté, une base de temps est déclenchée afin de contrôler le temps post-claquage.
Le mode de programmation avancé permet de contrôler de façon indépendante le courant et le temps poste claquage. L'étude de leur contribution sur le courant de lecture pourra donc être réalisée sur le plan mémoire de 1kb et par conséquent sur une grand population statistique. Des résultats quantitatifs doivent venir consolider l'étude présentée dans le chapitre 5.
16 démonstrateurs sont embarqués dans un véhicule de test. Une photographie est disponible à la figure 7.14 (p. 193). A l'heure de la rédaction du manuscrit de thèse, les tests n'avaient pas été réalisés en raison de planning très chargé des équipes compétentes. Les résultats seront éventuellement communiqués lors de la soutenance.

A.8 Conclusion

L'étude du claquage de d'oxyde ultra mince dans une gamme de haute tension n'étant pas couvert dans la littérature, un des objectif de cette est de mieux comprendre ce phénomène physique dans les gammes de tension de programmation. Un second de objectif est de porter des solutions au niveau systme afin de mieux exploiter le claquage d'oxyde de grille et d'améliorer les performances des mémoires antifusibles.
La caractérisation du claquage d'oxyde pendant la programmation d'une cellule nécessite des moyens expérimentaux dédiés. De nouvelles méthodes de test ont été mises au point au cours de cette thèse.
Celles-ci ont permis de mieux appréhender la physique du claquage. Un modèle de conduction Fowler-Nordheim s'est révélé pertinent pour modéliser le courant d'usure. Le "temps jusqu'au claquage" est quant à lui modélisé en fonction de la tension avec une loi de puissance. Ces deux modèles validés en haute tension ont permis de mettre au point une approche d'optimisation du dimensionnement de la cellule pour diminuer le "temps jusqu'au claquage" et la tension de programmation.

Un mécanisme de sur-courant de substrat a été mis en évidence par la caractérisation électrique pendant la phase de programmation. Le déclenchement d'une structure bipolaire PNP s'est avéré être une hypothèse pertinente. Ce mécanisme jusqu'álors non documenté n'est pas souhaitable du fait de la sur-consommation occasionnée. Une solution analogique a donc été implémentée pour limiter l'influence de cette structure parasite.
Une structure de test rassemblant une cellule élémentaire a donc été conçue pour permettre de contrôler le courant total circulant dans le point de claquage. Cette fonction permet donc d'étudier l'influence de l'amplitude de ce courant sur celle du courant de lecture et sur la caractéristique du point de claquage. L'étude quantitative résumée dans le chapitre 6 indique que le courant de lecture est en effet impacté par les conditions post-claquage de même que la partie résistive du point de claquage. Ce domaine de la physique du claquage d'oxyde mince n'est que très peu rapporté dans la littérature car il s'adresse à des applications spécifiques. C'es premiers résultats, bien que qualitatifs sont donc précieux et méritent d'être consolidés par une étude quantitative.
C'est la raison pour laquelle un démonstrateur a été conçu en technologie CMOS 32nm. Des fonctions avancées sont implémentées pour contrôler le courant et le temps post-claquage. La contribution de cette phase sur le courant de lecture peut donc être étudiée sur un plan mémoire de 1kb. La population statistique est beaucoup plus importante que celle couverte par des structures de test élémentaires. De plus, les cellules antifusible sont accédées dans un environnement de plan mémoire comme c'est le cas dans un produit industriel. Une second objectif de ce démonstrateur est de valider au niveau système, les fonctions avancées et d'en tirer parti pour optimiser les conditions de programmation.
Trois perspectives majeures sont identifiées :

- L'étude TCAD des mécanismes responsables du sur-courant de substrat mérite d'être confirmée. Comme ce phénomène est aussi présent sur la cellule antifusible fabriquée en 32nm, trouver une solution au niveau de la cellule élémentaire est important.

- Les tests sur le démonstrateur seront effectués prochainement. Il convient donc de vérifier la validité et la valeur ajoutée du mode de programmation avancé notamment dans l'optimisation de l'énergie de programmation.

- Les différents modèles proposés dans cette thèse méritent d'être confrontés aux technologies avancées. C'est le cas du procédé 32nm et du matériau diélectrique à haute permittivité. Cependant d'autres noeuds technologiques arriveront très prochainement avec d'autres innovations.

FOLIO ADMINISTRATIF

THESE SOUTENUE DEVANT L'INSTITUT NATIONAL DES SCIENCES APPLIQUEES DE LYON

NOM : Deloge DATE de SOUTENANCE : 15/12/2011

Prénoms : Matthieu, Yves, Claude

TITRE : Analysis of ultrathin gate-oxide breakdown mechanisms and applications to antifuse memories fabricated in advanced CMOS processes

Contribution à l'analyse des mécanismes de claquage d'oxyde ultra fin et applications aux mémoires antifusibles en technologies avancées

NATURE : Doctorat Numéro d'ordre : 2011-ISAL-0097

Ecole doctorale : Electronique, Electrotechnique, Automatique

Spécialité : Energie et systèmes

RESUME : Non-volatile one-time programmable memories are gaining an ever growing interest in embedded electronics. Chip ID, chip configuration or system repairing are among the numerous applications addressed by this type of semiconductor memories. In addition, the antifuse technology enables the storage of secured information with respect to cryptography or else.
The thesis focuses on the understanding of ultrathin gate-oxide breakdown physics that is involved in the programming of antifuse bitcells. The integration of advanced programming and detection schemes is also tackled in this thesis. The breakdown mechanisms in the dielectric material SiO_2 and high-K under a high electric field were studied. Dedicated experimental setups were needed in order to perform the characterization of antifuse bitcells under the conditions define in memory product. Typical time-to-breakdown values shorter than a micro second were identified. The latter measurements allowed the statistical study and the modeling of dielectric breakdown in a high voltage range, i.e. beyond the conventional range studied in reliability. The model presented in this Ph.D. thesis enables the optimization of the antifuse bitcell sizes according to a targeted mean time-to-breakdown value. A particular mechanism leading to a high bulk current overshoot occuring during the programming operation was highlighted. The study of this phenomenon was achieved using electrical characterizations and simulations. The triggering of a parasitic P-N-P bipolar transistor localized in the antifuse bitcell appeared as a relevant hypothesis.
The analysis of the impact of the programming conditions on the resulting read current measured under a low voltage was performed using analog test structures. The amplitude of the programming current was controlled in an augmented antifuse bitcell. The programming time is controlled by a programming detection system and a delay.
Finally, these solutions are to be validated using a 1-kb demonstrator yet designed and fabricated in a logic 32-nm CMOS process.

MOTS-CLES :Antifuse, dielectric, breakdown, non-volatile memory

Laboratoire (s) de recherche : Ampere UMR CNRS 5005

Directeur de thèse: Professeur Bruno Allard

Président de jury : Jean Michel Portal

Composition du jury :
Professeur Paolo Pavan
Professeur Gérard Ghibaudo
Professeur Bruno Allard
Philippe Candelier
Joël Damiens