SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di / Department of

Fisica G. Occhialini

Dottorato di Ricerca in / PhD program Fisica e Astronomia Ciclo / Cycle XXXIV

Curriculum Fisica Applicata ed Elettronica

One Time Programmable Anti-fuse Memory Based on Ultra-thin Oxide Breakdown : From Experiments to Test-chip Design

Cognome / Surname Gasparri Nome / Name Osvaldo

Matricola / Registration number 776549

Tutore / Tutor: Andrea Baschirotto

Supervisor: Paolo Del Croce

Coordinatore / Coordinator: Marta Calvi

ANNO ACCADEMICO / ACADEMIC YEAR 2020/2021

Questo documento conclude il mio percorso da studente. Un lungo ed impegnativo viaggio durato quasi 27 anni che ha visto protagonisti i miei genitori, debitamente vigili affinchè non mi facessi tentare da apparenze e semplici ma sbagliate strade.
Non so come abbiate fatto, ma ci siete riusciti:
Mamma, Papà, oggi sono più fiero che mai, fiero di voi e fiero di me.

Ad accompagnarmi lungo il percorso tanti amici e comparse, compagni di classe, colleghi. Grazie per esserci conosciuti, ho preso spunto da ognuno di voi per diventare chi sono oggi.

A voi, a cui non riuscivo a stare simpatico: grazie per avermelo fatto pesare. Ho vacillato sempre meno, con crescente dignità ed autostima.

Grazie ai maestri, professori e docenti. Grazie a Voi per le energie extra che talvolta ho richiesto. Grazie per le opportunità e preziosi insegnamenti extra-scolastici.

Grazie a chi ha condiviso con me quella gioia e ingenua spensieratezza che solo l'amore può dare. Grazie per aver conferito momenti di tregua alla mia perpetua razionalita'.

A mia sorella,
Per sempre grande conforto e sostegno.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The thesis retraces the research path aimed at finding the most suitable memory for trimming purpose in automotive applications. To begin with, an overview of existing memories is provided. Afterwards, the research will focus exclusively on the One Time Programmable (OTP) one. Finally, the functionality of the memory is described and the circuitry is designed and tested. The concept of a memory element is anything but new. OTP memory itself is actually a pretty old concept. There are many other more advanced memories that also allow more easy and fast programming: why dig into the past to recover something obsolete? Well, absolutely speaking, nothing is out of date, it all depends on the final application.

This introductory chapter first describes the most important characteristics of semiconductor memories and then identifies why, based on its characteristics, OTP is perfectly suited to the sought trimming application, the details of which will also be provided later.

1.1 MOS Memories

Metal-Oxide-Semiconductor memories can be divided into two main categories: Volatile Memories and Non-Volatile Memories.

In volatile memories, data can be stored by setting a state in a flip-flop bistable circuit: memories programmed with this approach are known as Static Random Access Memory (SRAM). A second programming method is based on charging a capacitor. However, the charge must be periodically updated due to the leakage current. Consequently, these are called Dynamic Random

Access Memory (DRAM). In both cases the information is lost if the power is removed: this is the reason why SRAM and DRAM are classified as volatile memories. Conversely, non-volatile memories (NVMs) are capable of storing data even when the power is off. The 95 % of the semiconductor memory market is occupied by DRAM and Flash, while the remaining 5 % by all other technologies. Volatile and non-volatile memories are manufactured as stand alone or embedded devices. Stand alone memories usually have a very high density. The most famous stand-alone memories are undoubtedly the DRAM and the USB smart key, using a NAND flash circuit.

A memory can also be incorporated into a System-on-Chip. In this case, data storage is not the main function of the circuit as for the new OTP applications.

OTP is one of the cheapest non-volatile memories (NVMs), even cheaper than EEPROM which requires additional processes. In fact, the conventional laser melting method has limitations with regards to laser wavelength, scalability and, most importantly, can only be programmed at the wafer level.

To make post-package programming possible, an electrically programmable NVM is required. The One-Time-Programmable (OTP), among the ROM memories, belong to the NVM family. There are many OTP bit-cell structures, each of which exploits various physical phenomena during programming and not all match the CMOS technology. Metal fuse, polyfuse and anti-fuse are the best candidates so far. Other memories require additional masks with respect to the main processes, additional technological steps or inaccessible programming conditions.

The challenges for the new OTP memories are: high density, high reliability, high memory capacity, low programming voltage, low programming current, speed comparable to that of flash memories and compatibility with the standard CMOS process for developing systems on chip, i.e. no additional mask for low-cost manufacturing. Such a memory does not yet exist.

OTP memories can be divided into: fused and anti-fuse. The difference is as follows: if programming causes the electrical conductive layer to become an open circuit, the component is called a "fuse". If, on the other hand, a layer acts as an insulator and the programming begins its electrical conduction, it will be called "anti-fuse". The operation of the fuse is mainly based on the melting of a conductive layer, while the operation of the anti-fuse is based on the breaking of the oxide.

Anti-fuse memories have been used for the past 10 years as embedded OTP memories for applications such as: code storage, secure encryption keys,

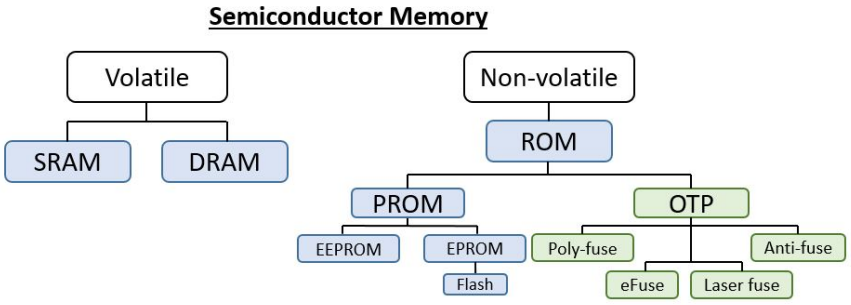

Figure 1.1: Semiconductor memories

chip ID, analog clipping and calibration due to their high level of security, low cost and non-volatility.

1.1.1 Random Access Memory

Random Access Memory (RAM) usually refers to volatile memories such as SRAM and DRAM. However, most NVM are organized to be randomly accessible: any bit of data can be accessed at any time. A simple block diagram of RAM is shown in Fig. 1.2. It consists of a memory array addressed via Word Lines (WL) and Bit Lines (BL). A memory bit-cell of any type is connected to each intersection. For example: an array of 32 WL and 32 BL contains cells at 1024 bits, with a memory density of 1024 bits (or 1 kb). The user selects which bit of the array to program or read. The selection takes place through a logic control block such as a WL and BL decoder, thus selecting a single WL / BL pair among all. In this way, only one bit-cell is accessed in the entire array.

In addition to the OTP, the ROM family also includes programmable ROMs, both electrically programmable (EPROM) and electrically erasable and programmable (EEPROM). USB memory, one of the best known, is an example of Flash memory, belonging to the EEPROM category. Flash memory has many interesting features: it is non-volatile, electrically programmable, electrically erasable and very dense. A flash bit-cell consists of a single device: a NMOS transistor characterized by a floating gate interposed between the channel and the typical gate oxide. The cross-sectional view of

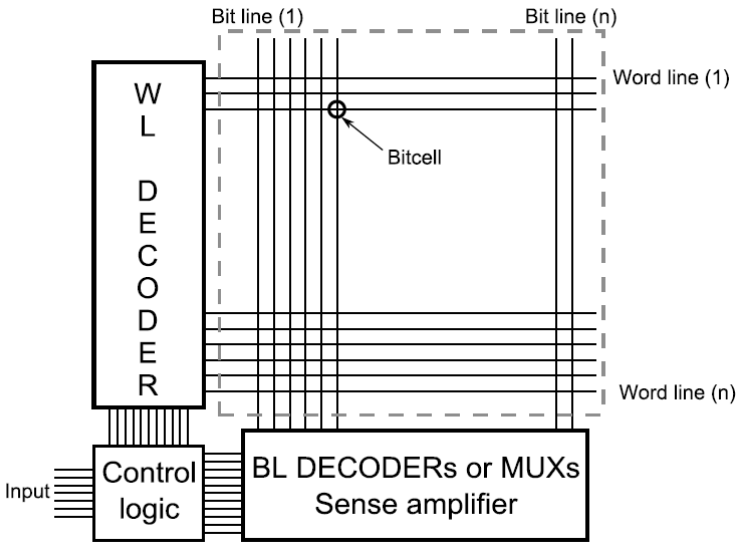

Figure 1.2: RAM block diagram

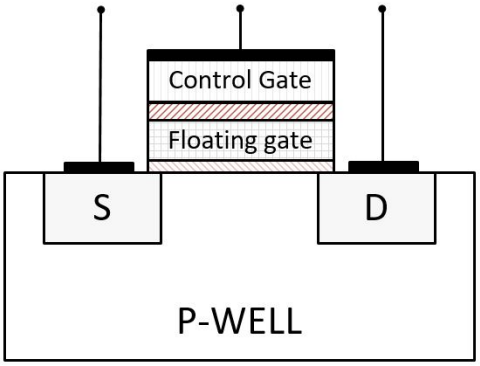

Figure 1.3: Flash bit-cell cross-section

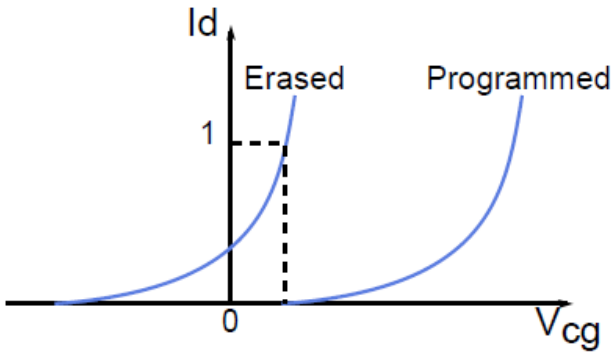

Figure 1.4: Flash bit-cell current VS gate voltage

a flash bit-cell is shown in Fig. 1.3.

The basic programming operation is to apply a high voltage such that the charges are trapped in the floating gate.

This results in a shift in the MOS threshold voltage. The cell content is related to the flash bit-cell current. A typical characteristic representing the drain current versus the control gate voltage is shown in Fig. 1.4. The floating gate is heavily insulated and trapped (negative) charges cannot be discharged for many years (under nominal conditions). To clear the flash bit-cell content, a negative V_{GD} is used to extract the electrons from the gate oxide. Then, the threshold voltage is reduced to its nominal value (minus trapped negative charge shielding the gate voltage needed to create the channel).

1.1.2 One Time Programmable (OTP)

Given their characteristics, none of the previous solutions are compatible with a logical CMOS process: they all require additional process steps and masking levels. In this context, the OTPs make their triumphant appearance.

As anticipated, the concept of one-time programmability is anything but new: it is about 65 years old. The storage matrix patented in 1957 was one of the first memories and was in fact one-time programmable. In fact, it was readable as many times as desired, but not re-programmable: here was the first name OTP or Programmable Read Only Memory (PROM). The device, or memory element, is intentionally stressed until it fails: a change occurs in its electrical conduction properties.

1.1.3 Laser fuse

The laser fuse is OTP's first integrated solution, reported in the 1980s. The fuse is a conductive element, such as a metal line. The memory is programmed by blowing the metal with a laser, thus creating an open circuit. Laser fuse technology is not easily scalable: very small bit-cells can lead to programming errors. Another limitation concerns the need for a laser itself, thus making the memory not electrically programmable. Of course, additional process steps, such as a higher level metal layer, are required to program the fuses using a laser machine. Also, due to the size of the laser spot (1.5 μm), metal links cannot be resized as desired.

All these features make laser fuse technology far from that sought after.

1.1.4 Poly-fuse

A poly-silicon line heats up to melting by means of a self-heating mechanism related to electrical conduction. The increase in resistance up to the final opening of the line is controlled by the conduction current. Localization of the melting point requires specific design of the poly-silicon link. The sensed current is compared with a reference current, the difference of which is then evaluated as a logic zero or one. While the material of the fuse melts with a high current, the phenomenon of electromigration occurs for a low current. Proper design of the poly-silicon link will favor one mechanism instead of the others.

Poly-fuses generally have short programming times ($\approx \mu s$) and low programming voltage. However, they require large programming currents (for example $10mA$) and need access transistors with large W to support them. Therefore they have a low density ($\#_{cell}/m^2$) at the bit-cells level (eg μm^2). An alternative to the poly-fuse is the metal-fuse: either placed on the same metal plane of a standard metal line or as an interconnection between two metal planes. Un-programmed metal link is a short circuit that will eventually open after programming. Although fuses are simple bit-cells, compatible with the standard CMOS process, the Cu requires a large programming current (and/or a high programming time) which leads to a large silicon area at the bit-cell level. Also, the bit-cells must be programmed one by one.

In the late 1990s, a new alternative to the laser fuse appeared on the market. The programming operation is based on the application of a high electrical stress to the bit-cell. Like a laser-fuse, a material is destroyed. Two

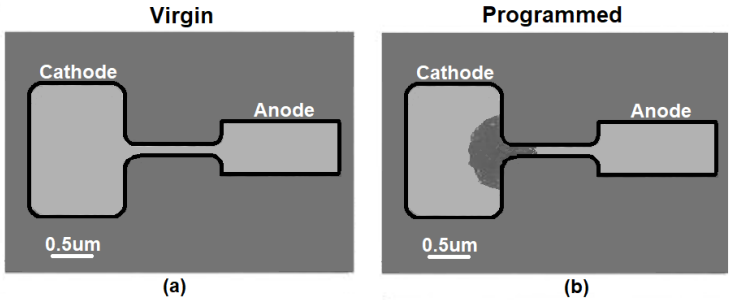

Figure 1.5: Virgin and programmed bit-cell

technologies are both widely used: eFuse and anti-fuse.

The programming principle of an eFuse is based on the explosion of a conductive element using a high current. In a short time its resistance increases as a consequence of the combustion of the conductive material. It can be a 2D poly-fuse, where a thin strip of poly-crystalline silicon is used as the fusible element or a 3D metal fuse which is a way between the second and third layer of metal. An example of a virgin and programmed poly-fuse bit-cell is shown in Fig. 1.5 (a): a thin strip of silicate poly-silicon connects the two electrodes (no further processes are required during manufacturing). The programming mechanism is based on electromigration: when a high current flows through the poly-fuse, the charges are mainly carried by the silicate layer due to its lower resistivity and eventually overheats. The depletion is located near the cathode (large electrode in Fig. 1.5 (b)). Consequently, in a programmed poly-fuse, electrons are carried by the poly-silicon resulting in a much higher resistance of the bit-cell.

The bit-cell with metal-fuse is similar to the poly-fuse. The interconnection via is placed between two metal layers giving a 3D structure. The area occupied by the metal-fuse is reduced compared to a planar structure such as poly-fuses. The programming mechanism is based on the metal electromigration: the conductor connection will be interrupted during programming.

Given today's replacement of the poly-silicon gate with the metal gate in CMOS technologies, the metal-fuse is becoming more popular than poly-fuses.

1.1.5 Anti-fuse memories

A capacitor can also be used as a memory element: it is nothing more than two metal plates separated by a dielectric. Therefore, when intact, the capacitor functions as an insulator, but in the event of a dielectric breakdown, the capacitor begins to conduct.

The programming mechanism consists in breaking down the capacitor by stressing the dielectric under a high voltage. On the atomic scale defects are generated in the dielectric until the formation of a contiguous path with low ohmic resistance that connects the two electrodes. The breakdown time of a capacitor is dictated by the size of the dielectric and the programming voltage: the thinner the gate oxide (or the higher the electric field), the shorter the breakdown time.

There are many bit-cell architectures. The basic concept is to have an oxide layer on top of a selector transistor (ST), a switch used to select which cell is to be programmed or read.

The most interesting bit-cell architecture are presented in the following section. The target cell should ensure:

- Adequate access transistor voltage tolerance
- An appreciable difference between the reading current of the programmed and non-programmed bit-cells
- The shortest programming time with reduced power consumption
- The smallest silicon area occupancy.

1T-1C anti-fuse

When a virgin cell is to be programmed, a high voltage is applied to the dielectric while the access transistor is ON (BL = HV and WL = Vdd). The dielectric is therefore stressed by a high electric field. Once the dielectric breaks down, the ST drain undergoes high voltage stress.

The dielectric layer is connected in series with an access transistor using thin or thick oxide technology Fig. 1.6. Its drain junction is subjected to severe stresses. In fact, the drain voltage is equal to $HV - R_{on} \cdot I$, where R_{on} is the resistance value of the programmed anti-fuse. The R_{on} value varies statistically with the programming condition and oxide size. This topic will be covered in more detail in Chapter 2.

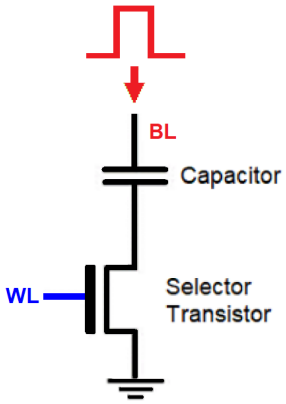

Figure 1.6: Architecture of a typical 1 transistor - 1 capacitor anti-fuse bit-cell

Instead of using a simple dielectric layer as an anti-fuse, a transistor can be used, where its gate oxide acts as a desired memory element. In this case the R_{on} will also depend on the breaking position (if it is near the source/drain region or away from it) and on the radius of the breaking point. The post-breakout characteristics are therefore not unique. In particular: if the BD spot is in the gate-surce or gate-drain overlap region, the gate current would behave ohmic. When instead it is in the channel region, the spot can be seen as an additional drain: the characteristic of the gate current would resemble that of the MOS transistor.

Due to past programming stress, the access transistor may fail or at least undergo premature aging. The reliability is critical. Also, in order to reduce the programming time, the voltage applied could be voltage significantly increased (however, still lower than the voltage capacity of the un-selected bit-cell: this defines the so-called *programming window*). For the thick oxide transistor the programming window is larger than thin oxide capacitor one. The larger the programming window, the more efficient the programming mode with a lower error rate, resulting in a more reliable bit-cell. A thick oxide transistor would be more robust, being less sensitive to high V_{GD}, but the density is in favor of the thin oxide transistor. In any case, the oxide of the transistor often remains stressed during programming mode. Another architecture comes into play to overcome this reliability limit.

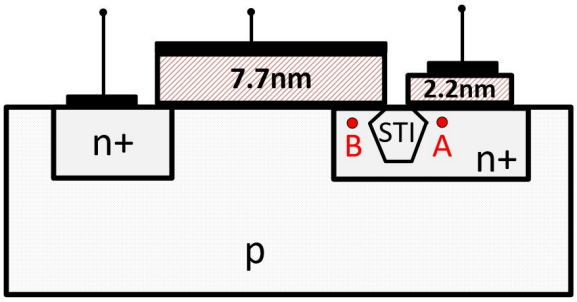

Figure 1.7: Cross-section of an anti-fuse bit-cell with a drift NMOS access transistor

Drift transistor anti-fuse

One solution to prevent stress on the drain of the access transistor is to use a drift transistor. The memory bit-cell comprises a dielectric layer and a thick oxide drift access transistor, as shown in Fig. 1.7. Such a cell generally uses additional Shallow Trench Insulations (STI) to be separated from neighboring cells, despite the area consumption. In the drift transistor, the conventional N+/Psub drain-to-bulk junction is replaced by a drain-to-Nwell junction. During the programming operation, the access transistor is selected and a high voltage induces the breakdown of the oxide. The voltage near the oxide interface (Fig. 1.7, position A) is abruptly shifted to the high voltage. Despite this, the voltage near the drain (Fig. 1.7, position B) is significantly lower. In fact, a voltage gradient is established inside the Nwell layer between points A and B, working as an internal resistance below the internal STI. This reduces the stress on the gate oxide of the transistor and increases the reliability of the bit-cells. The density of the bit-cells depends in part on the design rules: distances between L's, the wells etc. Furthermore, Nwell implants of adjacent bit-cells are subject to minimum spacing rules and this also limits the density of bit-cells. To increase the density, it is possible to use an anti-fuse bit-cell with drift PMOS access transistor. The Pwell layer plays the same role as the Nwell in the drift NMOS transistor, but the spacing rules are more favorable for the Pwell layer. However, the PMOS must have twice as much W/L as an NMOS for the same saturation current (due to the greater mobility of n carriers). In any case, no additional masks or process steps are required to manufacture the bit-cell drift.

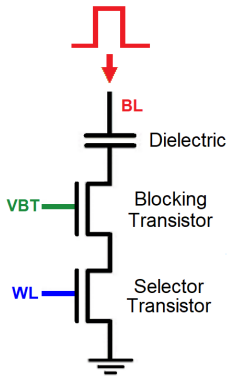

Figure 1.8: Cascode anti-fuse cell schematic

Cascode anti-fuse

A cascode access transistor could be used to improve the bit-cell reliability. A blocking high voltage (BT) transistor is inserted between the access transistor and the anti-fuse capacitor of the bit-cell 1T-1C as in Fig. 1.8. When the high voltage breaks the capacitor, the drain of the blocking transistor is exposed. However, its source voltage cannot exceed $VBT - V_{th}$, depending on the size of the device and the amplitude of the block voltage V_{BT}. This way the bottom transistor is well protected during programming mode. For example: 10V is applied to the BT drain after the programming event. $V_{BT} = HV - 4V$ with $V_{th} = 0.8V$, so its source will be 6V, while $WL = 3V$. The programming window will then be $> 4.8V$ and $< 11V$, otherwise the BT failure could occur.

1.2 Summary

Poly-fuse and anti-fuse appear to be the best candidates. Other memories require additional masks, additional technological steps or inaccessible programming conditions. Furthermore, the anti-fuse bit-cell is convenient for reasons of cell density and their low programming current allows to program an entire block of bit-cells, while the poly-fuse must be addressed one at a time. In any case, the programming voltage for the anti-fuse bit-cell must be large and properly controlled. The high programming voltage introduces reliability problems with respect to the access transistors. Furthermore, the

	Antifuse	eFuse
Process compatibility	⌣	⌣
Process variation	⌣	⌢
Programming voltage	⌢	–
Programming current	–	⌢
Programming time	⌢	–
Cell area	⌢	–
Complexity	⌢	–
Security	⌣	⌢

Table 1.1: Comparison between Fuse and Anti-fuse

programming time is quite large.

Although anti-fuse technology lags behind poly-fuse in terms of programming time, the gap narrows as the thickness of the dielectric shrinks. Despite a higher level of design complexity and slightly lower performance than eFuse memories, anti-fuse memories are the best solution for short-term product applications that require safety.

In 1T-1C bit-cells the programming current is limited by the access transistor which is always stressed during the programming mode. Of course, the cascode bit-cell is the most robust structure and has the highest programming windows, but it is still not the most convenient for area optimization.

A comparison of eFuse and anti-fuse properties is summarized in Table 1.1. The main disadvantage of the anti-fuse memory is the high amplitude of the programming voltage. Usually a charge pump circuit is used to generate such a high voltage. Through switched capacitors, stage by stage, the voltage rises up to six or seven times the nominal supply voltage. A poly-fuse or metal-fuse can be programmed using a normal supply voltage. However, they require a high programming current (e.g. $10mA$) while less than $1mA$ is sufficient to break the dielectric of an anti-fuse capacitor.

In the automotive field, with car battery-driven applications, the most suitable memory cell seems to be the anti-fuse one. Therefore, from now on, the study will focus exclusively on the anti-fuse cell. The 1T-1C architecture is the starting point of the investigation. In fact, apart from the preliminary dielectric characterization, the selector transistor is useful for selecting a cell among the many that make up the memory array: it serves both as a single-bit access and as an additional protection to preserve virgin, un-selected and

already programmed cell.

Cell performance strongly depends on the technology used. In fact, an HV selector transistor may already be sufficient to guarantee the reliability of the cell since it can withstand 30V of V_{DS} and 6V of V_{GS}, a condition outside the scope of the OTP.

1.3 IC Calibration and Trimming

After having considered the most important characteristics of the memories, it is time to describe the application that involves them.

Each Integrated Circuit (IC) takes some inputs and returns the desired output signals. An example in the automotive industry is the circuit behind the brake system. In a very intuitive way: the pressure on the pedal transforms into an analog signal and the brake caliper locks on the disc accordingly. Obviously all the same cars by default must have the same braking sensitivity and must be very precise. Furthermore, when the brake wears, the circuit must detect the variation or the same pressure on the pedal would result in less braking effect than in a new car. The brake system, the manufacturer choices, the car settings, all affect the analog circuits and their input and output signal specifications. Also, statistically, each chip may behave slightly differently due to the microscopical manufacturing difference. In fact, each transistor belonging to the same chip might behave slightly differently. The result would be a small and unpredictable deviation of the analog circuit and sensors from target specifications. The results is that two new and identical cars could actually therefore have very slightly different pedal sensitivity. The same would happen to the brightness of the on-board screen etc. Furthermore, the continuous downsizing of the transistor, aimed at achieving more ecological and economical solutions, inevitably increases the standard deviation sigma. The solution is called trimming: the displacement is detected and the references within the circuits update to reach that target specification.

From the design point of view, there are other techniques to deal with reference shifts such as chopping, auto-zeroing. However, these are not necessarily as accurate as desired due to side effects (charge injection, noise, mismatch, etc.). In fact, it is quite impossible to achieve maximum accuracy in analog signals even with the smartest and most expensive calibration algorithm. In fact, it is common practice for the chip of primary importance to

postpone any corrections of the references after the design phase. The test phase, in addition to ensuring the correct functioning of the circuit, allows to calibrate the circuit at the wafer level. This can be done in several ways. The basic concept is to measure output errors and subtract them from each input. However, to have a versatile circuit, able to adapt to different circumstances throughout its life without modifying it, this is still not enough. A further step is needed and this is where OTP memories triumph.

Suppose a reference depends on a resistance value which is statistically highly inaccurate due to manufacturing. One way to trim the circuit would be to include a series of different resistors connected to a metal-fuse in the design, then blow out any unnecessary connection until you reach the desired value. Somehow similar to trim the circuit with a potentiometer. Interesting, were it not for the blowing step, or programming phase, which requires new design/layout challenges: high current required, heat shielding for chip protection to perform. A better solution would be to use an OTP memory: higher memory density, lower power consumption, higher reliability. This without adding additional production costs. Embedded OTP contains a memory array in which clipping information is stored. Post-package programmability is fast and secure. Memory cannot be hacked. OTP is safer than eFuse: the latter allows the electron microscope to identify the melting point and read the contents of the memory. The anti-fuse based OTP, on the other hand, has as its programming method the dielectric break, the spot of which is so small that invasive means are required to read the memory bits.

Chapter 2

Wearout Current and T_{BD} Physical Models

Rather than actually using a capacitor or dielectric layer grown over the drain of the selector transistor, it is common practice to use gate oxide SiO_2 as an anti-fuse element. In this way, the metal gate would be the first armature of the capacitor and the source-drain terminals, in short circuit, would give the other. Indeed, gate oxide belongs to the main process of CMOS, it naturally adapts to the technology and does not require additional manufacturing efforts. Also, being very thin, it is the easiest oxide to break.

In the application under examination, an oxide layer was deposited in a window created specifically above the drain terminal ST as in Fig. 1.7. In fact, that is the memory cell concept chosen.

As anticipated, the programming mechanism of the anti-fuse OTP memory cell is the dielectric breakdown: a high voltage is applied to the ends of the anti-fuse element until its irremediable failure. The insulating properties of dielectric decay and an appreciable classical current begin to flow through it. Several charge transport processes can be involved such as direct tunneling for a low voltage or Fowler-Nordheim for a high voltage. A consequence of the carrier flow is the triggering of various physical mechanisms (release of hydrogen species, impact ionization, etc.). Therefore, numerous defects are created within the dielectric or at the interface. The accumulation of these defects leads to a conductive rupture path and dielectric rupture results. In general, there may be two different outs: hard and soft breakdown. When conduction no longer involves the tunnel effect, the dielectric breaks hard. For optimal programming, hard failure (physical disturbance of the

dielectric) is the only one allowed, otherwise it can be considered as a lack of programming.

This chapter explains step by step the theory behind dielectric breakdown: what happens when a dielectric is subjected to a high voltage stress. Together with the physical explanation of the degradation phases, equations are provided to predict the lifetime of the device, in order to characterize the dielectric breakdown as a function of the applied voltage, the size of the oxide, etc. The theoretical approach is eventually used to develop and implement OTP cells in the target 0.35 μm CMOS node. Usually about half of the memory area on silicon is occupied by peripheral circuits. One of the bulkier circuits is the charge pump to generate the programming high voltage. Therefore, understanding the influence of programming conditions is invaluable in optimizing surrounding circuits.

In reality this is not the case with the application under consideration as the required programming voltage is not that high and will be supplied directly by the car battery. In any case, the understanding of gate oxide breakdown is also of primary importance for improving the reliability of the cell and optimizing the programming phase.

There are different test methods for measuring failure parameters depending on how the stress voltage or the stress current are applied. In some breaking techniques, a relatively large number of capacitors are used to find the actual fault distribution, where the fault event is the anti-fuse programming. The cumulative number of faults, F, is depicted in a statistical graph. Then $ln(-ln(1-F))$ is plotted versus the programming field, linearly. Another technique is to apply a constant tension stress and measure the breakdown time. The latter, called Time Dependent Dielectric Breakdown (TDDB), sees $ln(-ln(1-F))$ plotted versus the programming time.

Forcing a current through an oxide layer can also cause failure. Either the current or the voltage is forced, the Time to BreakDown T_{BD} multiplied by the stress current I (or, equally, the HV multiplied by C_{ox}) is approximately constant. This leads to the introduction of the concept of the charge to breakdown Q_{BD}, defined as:

$$Q_{BD} = \int_0^{T_{BD}} I(t)\, dt \tag{2.1}$$

Generally for a good SiO_2 material, the charge density assumes values around $10 C/cm^2$.

A certain amount of charge, Q_{BD}, must be carried before a fault occurs. Since the dielectric is an insulator, with an extremely low concentration of free carriers, the charge carriers must be injected to reach Q_{BD}. Whatever electric field is applied, a current flows through the oxide layer, which ideally should be an insulator. What is this charge injection then? Where does it come from and why? The answers come from quantum mechanics: the tunnel effect. Electric charges have a non-zero probability of crossing a classically prohibited potential barrier. More specifically, two types of tunneling effects could arise: Fowler-Nordheim or direct tunneling. Of course, the greater the potential of the initial state, the greater the likelihood of overcoming a potential barrier. This is why an unsolicited dielectric is actually an excellent insulator: the probability of this injection is extremely low at nominal condition.

2.1 Statistical approach

The dielectric breakdown event must be statistically addressed by studying a series of anti-fuse induced failure to analyze the statistical dispersion of the event. To model a device or a system failure the Weibull distribution is commonly used [15],[16].

Percolation model

While the relevance of the Weibull statistic is well accepted, a model is needed that takes into account the different parameters of a capacitor. The so-called percolation model is based on the random generation of spherical defects within the dielectric material stressed under a constant voltage. Once a chain of defects is formed, causing a short-circuit of the two electrodes, the rupture event follows. This chain of defects is also called the percolation path.

The definition of the percolation model involves the diameter of a spherical defect a_0 (see in Fig. 2.1 (a)) or the mono-dimensional size of a ideal cubic cell broadly containing the defect (see in Fig. 2.1 (b)). The fault state is defined by a critical defects number N_{BD}, calculated as follows:

$$N_{BD} = \frac{t_{ox}}{{a_0}^3} \cdot A_{ox} \cdot exp\Big[- \frac{1}{\beta \cdot t_{ox}} ln\Big(\frac{A_{ox}}{{a_0}^2}\Big)\Big] \tag{2.2}$$

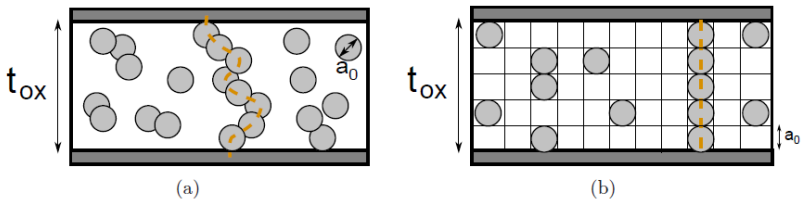

Figure 2.1: Representation of the percolation model

N_{BD} depends on the defect size a_0, the gate-oxide thickness t_{ox}, the capacitor area A_{ox} and the Weibull slope, β. In particular, β is defined as:

$$\beta \propto \frac{t_{ox}}{a_0} \tag{2.3}$$

The pertinence of the N_{BD} equation could be demonstrated as follows. Let's take a dielectric of the same size as a single cell of Fig. 2.1 (b). In this case $a_0 \cdot a_0 = A_{ox}$, which means that the exponential term in 2.2 is equal to 1. Also $t_{ox} = a_0$. As a result, the critical number of defects is unitary. Indeed, as expected, a single defect would already be enough to connect the two dielectric electrodes. Furthermore, looking at the dependencies, N_{BD} decreases as the longitudinal dimensions of the dielectric increase. In fact, the higher the oxide, the greater the likelihood of creating a percolation path. Also, the critical defect density should be achieved in a shorter time for a large device than for a short device due to a higher defect generation rate. This implies a short T_{BD} for larger devices.

2.1.1 Device Lifetime prediction

Experiments involving OTP memory are also used for the prediction of the transistors lifetime. In fact, the continuous scaling of the device made the gate oxide thinner and more delicate: the nominal and pn junction breakdown voltages of a given technology are shifted to lower amplitudes. Since the failure of a single transistor leads to the failure of the entire chip, a T_{BD} voltage-acceleration law is therefore crucial to ensure the reliability of electronic products for years.

The (HV, T_{BD}) couples data collected are fitted with empirical state-of-the-art voltage-acceleration laws, most of which are exponentially decreasing with the electric field magnitude or its inverse power. Then, the T_{BD} data

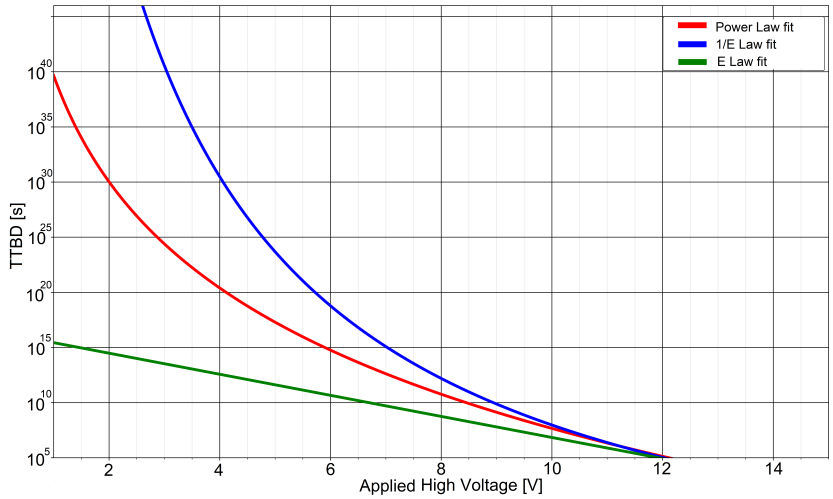

Figure 2.2: Comparison between mathematical laws

are interpolated down to nominal voltages and T_{BD} of years is extrapolated for the device lifetime prediction. Fig. 2.2 shows the three main relevant T_{BD} models, interpolating the data collected for a 7.7 nm gate-oxide in 35 nm CMOS technology. The models become more similar as the dielectric voltage increases, with an experimental time window enclosed in $10^5 s$.

Notation: the cathode is the electrode where a reduction semi-reaction takes place (loss of an electron / gain of a hole), the anode is instead the electrode where oxidation takes place.

2.2 Oxide Degradation Processes

2.2.1 Wearout current

Despite the insulating properties of an oxide layer, a leakage current flows whenever a voltage drop is applied. This phenomenon is amplified by reducing the thickness of the oxide as the electric field increases accordingly. The flowing current is called tunneling or wearout current. This quantum effect derives from the negative exponential solutions of the Schrödinger equation, which does not prevent the electron from crossing a potential barrier even if the electron's energy is less than the potential barrier (therefore unsurpassable for a classical particle). Two types of tunneling mechanisms can be

distinguished: Direct Tunneling (DT) or Fowler-Nordheim tunneling (FN). Their intensities depend on the applied electric field. However, they have a different formulation as the electron physically travels along different paths. In literature the threshold between the two is sometimes expressed from 8 to 13 MV/cm, but more generally between 5-20 MV/cm [3].

- For a low electric field, DT prevails. Here, the metal/oxide/Si band diagram is such that the electron crosses a trapezoidal potential barrier. The tunneling current is due to the electrons injected by the cathode which cross the oxide gap reaching the anode without flowing in its conduction band as in Fig. 2.3 (right).

  The approximate direct tunneling current, on the other hand, can be expressed using the same parameters A and B as follows:

$$J_{DT} = A \cdot {E_{ox}}^2 \cdot e^{-\frac{B}{E_{ox}} \cdot \left[1-\left(1-\frac{V_{ox}}{\phi_e}\right)^{3/2}\right]} \tag{2.4}$$

- As the field increases beyond the threshold, the current FN prevails: the field deforms the band diagram so that the electron now crosses a triangular potential barrier. Here, the electrons from the cathode reach the conduction band of the oxide before ending up in the anode. The illustration of the energy band diagram is shown in Fig. 2.3 (left). The formulation is given in Eq. .

$$J_{FN} = A \cdot {E_{ox}}^2 \cdot e^{-\frac{B}{E_{ox}}} \tag{2.5}$$

  Where A, B parameters are:

$$A = \frac{q^3}{8\pi h \phi_e} \cdot \frac{m^*_{Si}}{m^*_{ox}} \text{ and } B = \frac{8\pi\sqrt{2(m^*)_{ox}}\ {\phi_e}^{3/2}}{3hq} \tag{2.6}$$

  With q the elementary charge, (m^*_{Si}) the rest mass of the electron, m^*_{ox} the effective mass of the electron inside the dielectric and ϕ_b the height of the barrier of the injection electrode.

  Note that the current density for direct tunneling is higher than that of FN. Physically this is due to the neutral electron traps inside the oxide which help the hot electron to reach the anode through the so-called assisted tunneling trap [2].

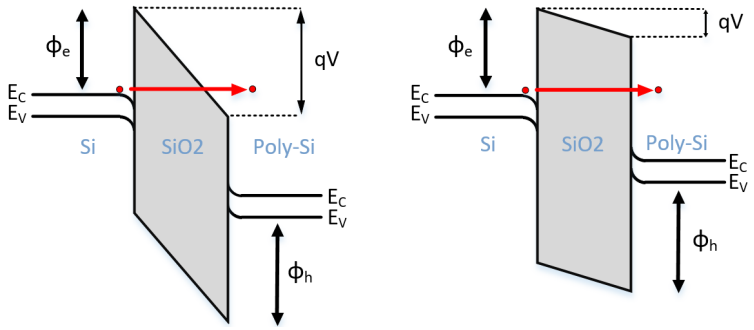

Figure 2.3: Energy band diagrams. Direct (left) and Fowler-Nordheim (right) Tunneling

For thick oxide ($t_{ox} > 5\,\text{nm}$) and electric fields above $5-8MV/cm$ the FN current dominates [1]. While for ultra-thin oxide ($t_{ox} < 5\,\text{nm}$) and voltage below $3.1-3.2V$ (corresponding to the typical height of the potential barrier between n-doped silicon and the gate oxide) direct tunneling heads. In both mechanisms the transport is limited by the height of the electrodes which are ϕ_e for the electrons and ϕ_h for the holes. Since the height of the conduction band barrier, ϕ_e, is less than the height of the valence band barrier, ϕ_h, conduction is expected to be dominated by negative carriers.

There is actually a third transport mechanism, called the Frenkel-Poole transport. In that case, previously trapped electrons are emitted in the conduction band of the oxide. This mechanism is triggered by temperature: in fact, a charge might move from one trap to another due to thermal excitation. In this case the conduction limitation is not the height of the electrode barrier, but rather the height of the trap barrier. This mechanism is critical when both the temperature and the electric field are high, in a dielectric with a certain amount of traps.

$$J_{FP} = C \cdot E_{ox} \cdot e^{-\frac{q\phi_t}{kT}} \cdot e^{\frac{\beta_{FP}\sqrt{E_{ox}}}{K_b T}} \tag{2.7}$$

Where C depends on the defect density and β_{FP} is the Frenkel-Poole factor defined as:

$$\beta_{FP} = \sqrt{\frac{q^3}{\pi \epsilon_{ox}}} \tag{2.8}$$

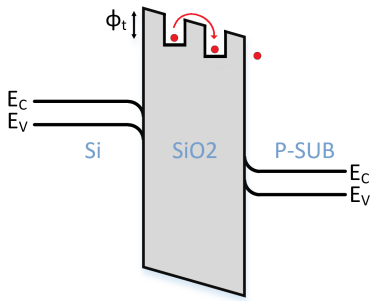

Figure 2.4: Energy band diagram: Frenkel-Poole Tunneling

For a given E_{ox}, the current density DT is greater than that of FN. Based on the tunneling regime, differences in the Breakdown event can be expected, such as the time at which it occurs. The electric field threshold (E_{ox}) between DT and FN more generally can be considered between $5 - 20\,\mathrm{MV/cm}$, for example for $t_{ox} = 7.7\,\mathrm{nm}$ the corresponding electric field would be $\simeq 3.8\ rmV$. Since $E_{ox} = HV/t_{ox}$, where HV is the voltage applied across the dielectric, any variation of t_{ox} would affect the MV/cm ratio. In the construction of an OTP cell with 350 nm technology ($t_{ox} = 7.7nm$, the same used for the drift transistor gate oxide) in the high voltage domain ($> 10V$), the $E_{ox} \simeq 13\,\mathrm{MV/cm}$, so the FN current should dominate.

2.2.2 Defect Generation Mechanisms

Defects arise from imperfections in the dielectric lattice such as the presence of foreign atoms or from generation mechanisms. The defect could introduce an energy level into the forbidden SiO2 band-gap that can act as a trap or recombination center for those carriers tunneling the oxide. For this reason they are simply called traps. Energy levels can be distinguished according to their depth. If the link between the carrier and the defect is weak, the level is not deep since it is close to the conduction band. While if the bond is strong it means that the level is very deep, as it is very far from the conduction / valence band. When electrons are funneled into the oxide, they trigger different defect-making mechanisms depending on their energies. There are many possibilities of creating defects with different impact on the oxide. In particular, the electrons involved in the scattering events and in the breaking of the bonds are mainly responsible for the degradation of the oxide. First, electrons can be trapped in pre-existing SiO2 traps (eg oxygen

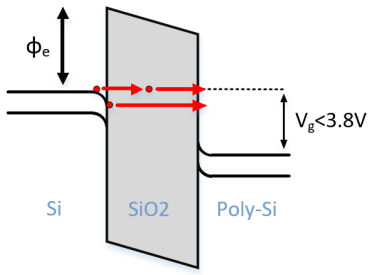

Figure 2.5: Trap assisted tunneling in DT (interface and deep traps)

vacancies, generally due to the presence of OH or H2O). Likewise, holes can be trapped but due to their less mobility they hardly break free. However, these mechanisms do not play a large part in oxide degradation. The most important traps are those inside the oxide or at its interfaces, generated by hot electrons with energy higher than 2eV (referred to the bottom of the oxide conduction band). These electrons can release hydrogen from the defect sites near the anode interface. These hydrogen-like species then travel to the oxide reaching the cathode interface, where they produce interface states.

In a trap-assisted tunneling, electrons pass from the cathode to the trap and then from the trap to the anode, as shown in Fig. 2.4 (left) for a deep (top) and an interface trap (bottom).

For electrons with energy $> 5\,\text{eV}$ (as in the case of the higher voltages required in OTP Breakdown) the Anode Hole Injection (AHI) mechanism and the bonds rupture, shown in Fig. 2.4 (right), are activated [1], [2]. The hot electrons channel the triangular barrier reaching the conduction band of the anode. Then, they transfer their energy to the deep valence band electrons, which are promoted to the conduction band leaving holes. Due to E_{ox}, the hot holes are channeled into the valence band of the oxide. This could lead to the generation of interface traps and recombination centers for the incoming electrons [8]. This is followed by trap-assisted tunneling at localized points, leading to irreversible oxide damage.

This means that the electrons have already lost their energy along the way. However, with ultra-thin oxide ($< 5nm$) the dispersion is considered negligible: to a good approximation the transport is ballistic. Therefore, the electrons reach the anode conduction band where they give up their energy to generate holes for scattering ionization. In fact, electrons elastically transfer all their energy to an electron in the deep valence band. The latter

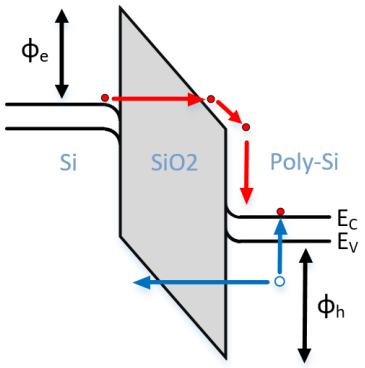

Figure 2.6: AHI mechanism

will then be promoted to the lowest available state: the anode conduction band, leaving a hole in the valence band. According to the direction of the electric field, the hot hole returns to the oxide layer in its valence band, with a speed lower than that of the electrons. In part, they can be trapped in deep sites (e.g. an oxygen vacancy) of the oxide at the cathode interface or release hydrogen from Si-H or SiO-H bonds. Therefore, subsequently injected electrons that tunnel into the metal / oxide potential barrier can recombine with trapped holes producing interface states and traps near the cathode. Furthermore, the mechanism leads to an increase in current density in localized spots due to the generation of traps followed by trap-assisted tunnels, leading to irreparable oxide damage. This mechanism is highly dependent on the thickness of the oxide for the ultra-thin oxide.

Hydrogen release dominates in that voltage range in which MOSFETs operate, making this more attractive for device life prediction studies. While impact ionization and anode hole injection occur for higher voltages and therefore are of primary importance in OTP characterizations. The energy thresholds between the mechanisms are: 2eV hydrogen release, 5eV anode hole injection and 9eV impact ionization [1]. Note that a given probability of defect generation depends on the size of the oxide. Indeed, considering the basic electric field for a capacitor with two parallel plates of surface S and separated by a distance d (by Gauss's law):

$$\overrightarrow{E} = \frac{\sigma}{\epsilon} = \frac{\Delta Q}{S \cdot \epsilon} = \frac{V}{d} \tag{2.9}$$

It is clear how a change in S and d would affect the MV / cm ratio used to distinguish the defect generation mechanisms. In general, the defect generation ratio is expected to be a function of electric field, lattice imperfections, humidity, pressure and temperature.

2.3 Breakdown Physical Models

The efficiency and reliability of the programmability of the OTP depend on the accuracy of the evaluation of T_{BD}, the value of which depends on the dominant mechanisms of creation of the defect. There are three main mechanisms that produce Breakdowns, each of which leads to a different T_{BD}.

2.3.1 E Model

Also known as the Thermo-Chemical model. This model suggests that the break is the result of the effect the electric field has on atomic bonds. In other words, the generation of the defect is a field-driven process, while the hot electrons passing through the oxide can be considered a secondary effect. This is due to the polarity of the SiO_2 bond, where electrons are centered in O atoms leading to an electric dipole composed of positively charged Si ions and negatively charged O ions. Due to an external electric field, the covalent bond in the oxide degrades until the dipoles break as a result of the electrostatic interaction. In this case:

$$T_{BD} = C \cdot e^{-G \cdot E_{ox}} \cdot e^{\frac{E_a}{K_b T}} \tag{2.10}$$

Where G is the electric field acceleration factor and E_a is the activation energy for the oxide breakdown.

2.3.2 1/E Model

Also called the Anode Hole Injection (AHI) pattern because it suggests that the break comes from the hot holes injected by the anode. The tunneling current of the hole can be expressed as the product of the tunneling current FN of the electron and a term that expresses the probability of the generation of the hole and of the tunneling through the anodic barrier (which also has an $exp(-1/E)$ behavior, such that their product still scales exponentially with

$1/E$ [4]). The total charge of the holes, Q_P, can be expressed as the integral of the tunneling current of the hole in the time window $[0, T_{BD}]$. Due to the exponential dependence of the current FN on the applied voltage, even the breakdown charge, $Q_{BD} = Q_P/J_{FN}$ reflects the same behavior leading to a Q_{BD} which decreases when the voltage applied to the oxide increases [4]. In particular, assuming that electronic tunneling is described by Eq. 2.5, the time to breakdown can be obtained as: $T_{BD} \simeq Q_{BD}/J_{FN}$ (not identity because traps can lead to small temporary changes in the J_{FN}). The $TTBD$ is then obtained from the equation:

$$T_{BD} = D \cdot e^{\frac{F}{E_{ox}}} \cdot e^{\frac{E_a}{K_b T}} \tag{2.11}$$

2.3.3 Power Law

The power law model is related to the hydrogen release phenomena at the dielectric interface. Indeed, the hot tunneling electron can break the Si-H bonds, leading to the release of H atoms at the metal / oxide interface. These can subsequently diffuse through the dielectric layer and combine with oxygen vacancies. In this way, defects are generated until the breakdown occurs. The equation is therefore:

$$T_{BD,Power} = K \cdot {V_{cap}}^{-\beta} \tag{2.12}$$

where β is the voltage acceleration factor, correlated to the breaking energy of the Si-H bonds [5].

2.4 TDDB modeling

The time-dependent dielectric breakdown (TDDB) model aims to study the dependence of the breakdown time (T_{BD}) on the stress voltage.

The best product-like OTP memory would be characterized by short T_{BD} and low HV, meaning: minimum programming energy with maximum failure effectiveness. In general, the lower the programming power, the easier is the effective integration of the charge pump circuit involved in generating the programming voltage, with less occupation of the silicon area even if this is not the case for the battery powered application.

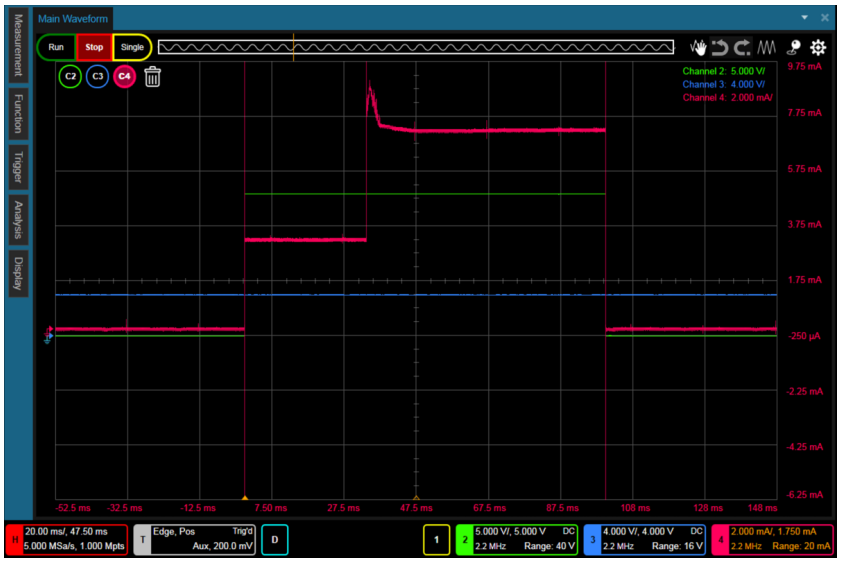

Figure 2.7: Breakdown event seen on the oscilloscope

An example of T_{BD} measurement is shown in Fig. 2.7 for a cell with $T_{ox} = 7.7nm$ belonging to the Infineon Technologies HV transistor technology. The fault is clearly distinguishable due to a sudden current flow. After the application of the HV pulse (in green), a wearout current flows. After a lapse of time T_{BD}, the current suddenly increases until it settles at the value of the saturation current of the selector transistor. During the wearout phase, the detected current is attributable to the tunneling effects and is of less amplitude than the ST saturation current. It is safe to assume that the ST is working on its linear regime during the wear phase. This result will be used for simulation and modeling of the bit cells during the programming phase. In fact, the drain-to-source ST voltage must be evaluated and taken into consideration to accurately determine the voltage across the capacitor.

This can be done by plotting the wearout current characteristic of the anti-fuse element and the IDrain - VDrain characteristic of the drift transistor so that the duty point is determined at the intersection of the two curves.

As anticipated, most of the dielectric characteristics could be deduced from the analysis of the wearout data. This can be done by applying a voltage ramp to the dielectric electrodes. For each value, the current is listed. However, the equipment usually loses in accuracy when it comes to fast transactions ($\simeq ns$) as it happens close to the BD point. To overcome

this problem, missing data could be extrapolated.

2.5 Wearout current modeling

A model of the leakage current through a dielectric is presented here. A Fowler-Nordheim tunneling mechanism can be hypothesized due to the high amplitude of the programming voltage.

The DC and transient measurement data are reported in Fig. **??** as the current density J_{cap} with respect to the electric field E_{cap}. Both parameters are calculated as:

$$J_{cap} = \frac{I_{cap}}{A_{cap}} \quad and \quad E_{cap} = \frac{V_{cap}}{t_{ox}} \tag{2.13}$$

The calculation of the electric field through the dielectric, E_{cap}, is given by the following equation:

$$E_{cap} = \frac{HV - V_{FB} - R_{on} \cdot I_{wearout}}{t_{ox}} \tag{2.14}$$

where V_{cap} is the voltage applied to the capacitor, V_{FB} is the flat band voltage, R_{on} is the ST on-resistance and I_{cap} is the tunneling current. Since the drift transistor operates in a linear regime during the wearout phase, the on-state resistance (Ron) of the selector MOS device is:

$$R_{on} = \frac{L}{\mu_n C_{ox} W (V_{gs} - V_{th})} \tag{2.15}$$

In the case of the anti-fuse drift bit cell, the term $R_on \cdot I_{cap}$ must be taken into account in correcting the voltage drop in the drift transistor. The flat-band voltage, on the other hand, is negligible ($\simeq 1V$) in a first approach [14].

For example: considering an oxide thickness of $2.2nm$, a HV = $8V$, a wear current of $150\mu A$ with a resistance of $1k\Omega$ for the selector transistor in wearout phase, from (2.14) we obtain an $E_{cap} = 6.75MV/cm$. Thus, theoretically, FN conduction is expected to describe the phenomena. However, it must be considered that the thickness of the oxide is very small and usually when it is less than 5 nm, direct tunneling will become the dominant conduction mode. For this reason, in reality, when fitting the data with an FN model, discrepancies are expected because in reality the correct model

would have to deal with both loss mechanisms. However, it is noted that the correct fit in the figure appears for an electric field greater than 25MV/cm. This is due to parameters neglected in the calculation of the effective electric field applied across the capacitor, such as the flat band voltage.

Fowler-Nordheim conduction modes are driven by bias voltage and electrode barrier height, temperature dependence is negligible, while trap-assisted mechanisms such as Frenkel-pool transport are strongly activated by this parameter. Therefore it could be interesting to evaluate the response of the $I(V_{cap})$ system as the temperature varies. How temperature affects the breaking point is much more interesting from an OTP point of view and will be discussed later.

2.6 Optimization of anti-fuse cell

Now the aim is to minimize the amplitude of the programming voltage, the break time and the size of the anti-fuse bit-cell. A circuit equivalent to an anti-fuse bit cell during the wear phase is proposed to derive an equation of the capacitor voltage as a function of the programming voltage, the area of the capacitor and the dimensions of the drift transistor. Hence, it is possible to study the impact of these parameters on breakdown time and wearout current.

In order to calculate the programming voltage required to have a reliable anti-fuse bit-cell failure, the procedure is the following: - First a target time-to-breakdown must first defined, e.g $10\mu s$
- By knowing the area of the anti-fuse capacitor, the empirical T_{BD} law is used to deduce the V_{cap} voltage, e.g a power law (2.12) with $n \simeq 43$ for a $1\mu^2$ capacitor
- (2.15) is used to obtain the R_{on}
- The current flowing, which would follow the FN formulation (2.5) is then used to calculate the required HV

$$HV = V_{cap} + R_{on} \cdot I_{FN} \tag{2.16}$$

Now, the amplitude of the programming voltage calculated using the presented algorithm can be plotted as a function of both the area of the capacitor and the width of the transistor. The minimum value can be obtained for a better efficiency / cost of the bit-cell. The minimum HV should be obtained for the maximum width of the drift transistor (i.e. the smaller R_{on} since the

lower the state resistance, the lower the voltage drop across the access transistor and therefore the greater V_{cap} for a given HV) and approximately the minimum size of the capacitor. Although a large capacitor shows a shorter T_{BD}, the voltage acceleration term (V_{cap}^{n}, n = -43) is more significant than the proportionality factor (1 / A_{cap}). This is why the wear current contribution is the dominant factor.

However, there are two limitations to using a large drift transistor. First, there is an obvious penalty in density, in disagreement with the growing demand for dense anti-fuse memories, the reduction in capacitor area is obviously valuable. Second, a large channel leads to a high post-breakout current. While sufficient post-programming current is required to activate a satisfactory reading current, it has a direct impact on power consumption.

2.7 Models Validation

Based on the physical mechanisms described above, the case of an OTP cell in technology 350 nm with an oxide surface of 9.2 μm^2 and a thickness of 7.7 nm is studied. The goal is to relate the electric field, the oxide properties and T_{BD}. To do this, high voltage pulses (> 13V) are applied to the gate (while the source and drain have been shorted) to induce the breakdown. From Fig. 2.8, for a 13 V programming voltage, the wear current is approximately 50 μA and T_{BD} is 30 ms. Since the electric field is > 15 MV/cm, the current FN dominates, as shown in Figure 2.9 where Eq. 2.5 fits the data correctly. Each data is the average of 5 samples. Despite the low statistic, the relevance of the FN behavior is still verified. The extrapolated parameters A and B of the eq. 2.5 are:

$$A = 3.5 \cdot 10^{-3}\,\mathrm{A/V^2} \quad \text{and} \quad B = 3.8 \cdot 10^{10}\,\mathrm{V/m} \tag{2.17}$$

The behavior of T_{BD} as a function of the applied voltage is shown in Fig. 2.10. The fitting equations are 4.11, 2.11 and 2.12 where $V_g/t_{ox} = E_{ox}$. The extrapolated parameters are:

$$\begin{aligned} E\ model:&\ (C \cdot e^{\frac{E_a}{K_b \cdot T}}) = 1.2e^{11},\ \ G/t_{ox} = 2.22 \\ 1/E\ model:&\ (D \cdot e^{\frac{E_a}{K_b \cdot T}}) = 2.4e^{-17},\ \ F \cdot t_{ox} = 460 \\ PowerLaw\ model:&\ K = 2.3e^{34},\ \ \beta = 32 \end{aligned} \tag{2.18}$$

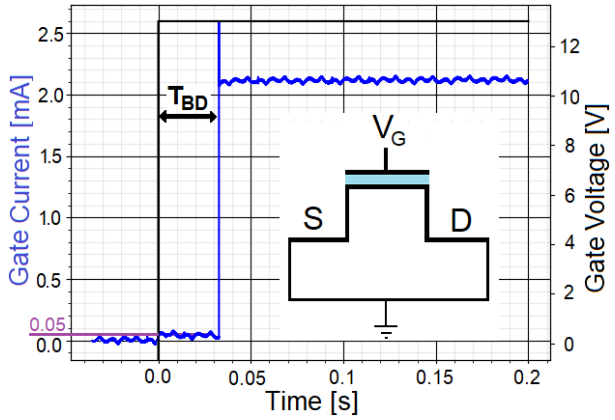

Figure 2.8: Gate current waveform for T_{BD} detection with a 13V V_G for a CMOS anti-fuse element

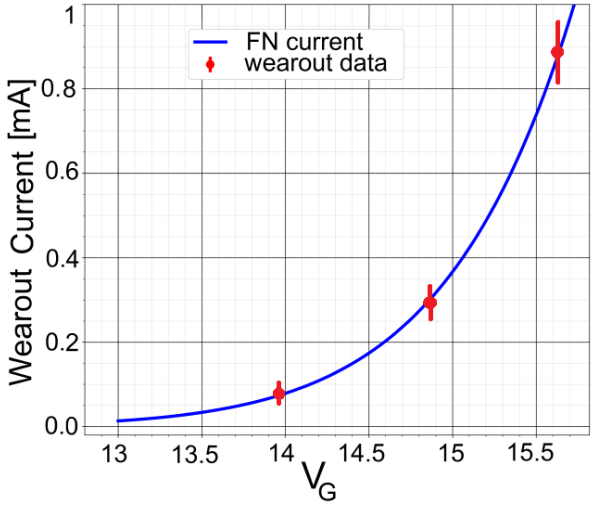

Figure 2.9: Wearout current fit with FN equation

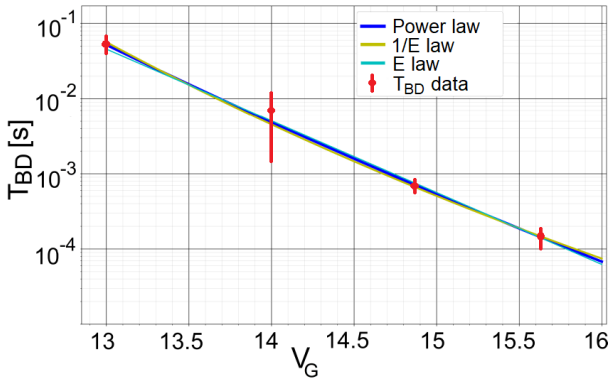

Figure 2.10: T_{BD} data fit

These results are also validated by the good agreement with the literature. In fact, B is consistent and the parameters of model E are comparable with those obtained respectively in [6],[7].

Adaptation to experimental data demonstrates the pertinence of the three models. However, the experiment is still not sufficient to determine which of these defect generation mechanisms has the greatest impact. In fact, the breakdown is generally a consequence of multiple overlapping mechanisms. However, the assumptions of T_{BD} can still be obtained by extrapolation. It should be noted that it is important to guarantee a maximum failure rate: the programmed oxide must be hardly broken down, to avoid misunderstandings when reading the memory. The greater the electric field, the greater the likelihood of having hard breakdown in the programming time window. However, producing a higher voltage would require more power and area consumption, affecting the overall cost of the chip. On the basis of these considerations, the specifications regarding the oxide properties, the target T_{BD} and the applied field are drawn and the design of the OTP circuit follows.

2.8 Conclusion

The breakdown of the oxide is the final, irreversible stage, caused mainly by hot electrons due to the applied electric field. The models presented here can be used for duration estimation experiments. In particular, once the voltage drop across the oxide is known, the TTBD equations can be used to predict the worst case lifetime of the device. Otherwise, in OTP

memories, starting from the desired TBD, the procedure could be reversed to optimize the cell in terms of power and size. In general, when validating models, exponential time-to-breakdown behavior should not be regarded as the triumph of a particular physical model. In fact, as shown in Figure 4, the difference between the three models is not very noticeable. As explained, there is no single mechanism for generating the defect, but rather it is an overlap of all the mechanisms involved.

Chapter 3

Set-up and Experiments

The first step, before the design of the OTP cell, is the careful characterization of each component of the bit-cell. Not only the dielectric must be studied in depth, but also the selector transistor, which plays a crucial rule in the functioning of the bit-cells. In fact, since the anti-fuse layer will be made by growing the oxide on the drain of the selector transistor it is more convenient to start the characterization with stand alone selector transistors. Finally, the next step would be to study the selector transistor with the implanted anti-fuse element.

Wear current and T_{BD} data will be collected by varying the ST and dielectric dimensions (area and thickness), applied voltage and temperature. The data will be analyzed and processed with specially written Phyton scripts (not covered in the document). Fig. 3.1 shows the experimental procedure performed for both ST and dielectric characterization. The wearout current sensing configuration includes a high voltage pulse generator and an oscilloscope with probes for collecting current data. The waveform of the bit line current is represented using the 50Ω oscilloscope inputs. Due to the very rapid $\simeq nS$ transient breakdown event, some details during acquisition may be lost due to the setting's inherent maximum sample rate. Although very basic, the above setting allowed to collect the searched data relating to:

- the Time-to-BreakDown (T_{BD}): the x value of the current transient of the bit line in which the positive step occurs;
- the wearout current: the current y value between the application of the HV pulse and the T_{DB};

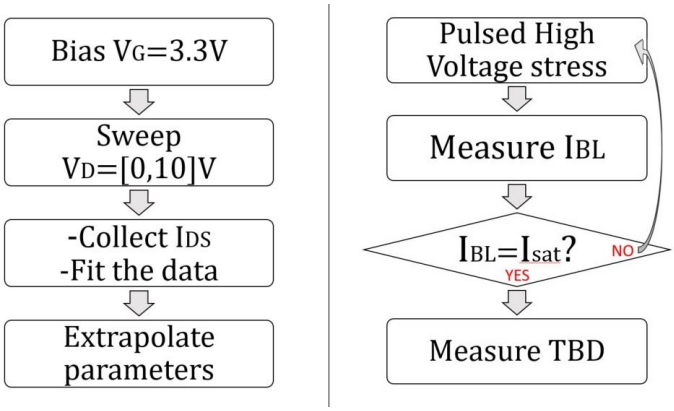

Figure 3.1: Characterization flowcharts: Selector Transistor (left), Dielectric (right)

- the post-break current;

The above data will be interpolated with the empirical physical models presented in Chapter 2, to obtain information regarding:

- the transport mechanism and defects: injection of holes, creation of traps;
- the formation of low-ohmic resistance paths;
- the behavior of the broken down capacitor.

As in Fig.3.2, Bulk and Source are short-circuited. When measuring the current flowing in the Drain terminal, the gate current of the access transistor being considered negligible, the probe detects the bit-line current: $I_{BL} = I_D + I_{bulk}$. During the wearout phase, the bulk current is negligible. When approaching the rupture event, however, the bulk current contribution is significantly higher [17].

The experiments are conducted at the silicon wafer level. The technology department has prepared some pad-rows for the preliminary characterization of the OTP: some with the anti-fuse element and others without, to test exclusively the selector transistor (also made specifically, without any cadence model or data sheet parameters). Many separate bit-cells are placed on each row, which change the size and ST drift regions. Also, for statistics collection, some pad-rows contain arrays of the same bit-cell. In this way, multiple cells are programmed simultaneously with a single HV pulse.

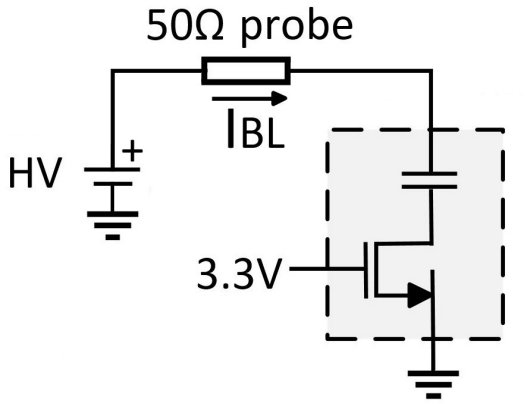

Figure 3.2: Experimental set-up for transient programming current measurements

Subsequently a second silicon wafer was produced to study also the anti-fuse with an oxide thickness of 2.2 nm, in addition to the nominal one of 7.7 nm (keeping the selector transistors unchanged). The wafers are tested using a pico-probe station with 4 needles to polarize the bit cell terminals and extract the signals of the wafer reticle under investigation. The HV pad, however, is not directly connected to the source generator. The signal, in fact, coming from a waveform generator, must first be amplified and then applied to the HV pad. This is used for both ST and dielectric characterizations. In the first case, a triangular waveform is used as a sweep on the drain voltage ST; in the latter, instead, a quadratic impulse is applied to induce rupture.

Afterwards, to improve the accuracy of the set-up, the pulse programming application was changed after a few measurements. In fact, the current transient showed peaks of $\simeq \mu s$ with each pulse application. This is a consequence of the parasite coupling effect, due to the fact that the current sense also insists on the same HV node. This would be a major limitation in TT_{BD} experiments, where $\simeq ns$ precision is actually required (and already difficult to achieve). Otherwise, the spikes could obscure the wear current and valuable T_{BD} data. The programming pulse is actually applied to the ST gate. The BL continues to be DC biased with HV, but only when a 3.3V pulse is applied to the ST gate the current begins to flow (of course, the high voltage ST device must be able to withstand that HV across the gate and drain terminals). This update avoids problems when reading the BL current.

Due to the 50 ohm current probe, for a more precise calculation, in any calculation, however small, the voltage drop across it should also be considered. For the larger device (#4), the voltage drop will be: $6mA \cdot 50\Omega = 300mV$, thus slightly varying the voltage applied to the anti-fuse element. The effect becomes more evident and no longer negligible when dealing with a thinner 2.2nm oxide.

3.1 Selector transistor

As anticipated, to ensure the reliability of the bit cell, the ST is a high voltage device, having a gate oxide thickness of 7.7 nm and a drift area under its drain to lower the electric field between the HV node and the gate. Since it was produced ad hoc for the OTP application, it has never been modeled, only layouted. In particular, from the point of view of the layout, they are almost identical to a complete ST + anti-fuse bit cell. The only missing processes concern the deposition of oxide on the visible oxide windows, left empty. Not only is the spike model not yet available to simulate its response, but its precise parameters, such as R_{on}, are also missing and must be computed. To do this, it is essential to obtain the output characteristic of the device, mainly to extrapolate the resistance of the triode region and the saturation current. These results are later used in the bit-cell experiments to calculate the actual voltage drop across the anti-fuse element, during the wearout phase, and detect the hard breakdown event. This can be seen as the time interval in which the ST suddenly changes its operating region from linear to saturation.

By increasing the HV node with a voltage from 0 to 12 V, biasing the gate to 3.3 V, the current data are stored and the I-V characteristic is obtained. A Python script does the rest of the work: plot the I-V data and fit it into the triode region equation (3.1) to extrapolate the μC_{ox} parameter and the linear resistor.

$$I_{DS} = \mu C_{ox} \frac{W}{L} \left((V_{gs} - V_{th}) V_{ds} - \frac{V_{ds}^2}{2} \right) \tag{3.1}$$

The same procedure was repeated for 12 transistors: all having the same length $L = 0.45\mu m$, but different width W and length of the drift region. In particular, half of them have a long drift region and the other 6 have a short

Device	W	L	W_{ox}	L_{ox}
1	5.1	0.45	1.5	0.8
2	7.1	0.45	3.5	0.8
3	9.1	0.45	5.5	0.8
4	15.1	0.45	11.5	0.8
5	5.1	0.45	1.5	2
6	5.1	0.45	1.5	4

Table 3.1: Devices dimension (in μm)

drift region. In reality, the devices #1, #5, #6 are identical: the difference lies in the size of the oxide windows, therefore appreciable only if the oxide had been added.

Table 3.1 summarizes the dimensions of the transistor (valid for both the long and short drift region), including those of the oxide windows, W_{ox} and L_{ox}.

3.1.1 Results

Fig. 3.3 provides an example of I-V fitting (device #1) of the collected data (in red) with the equation (3.1) (in blue).

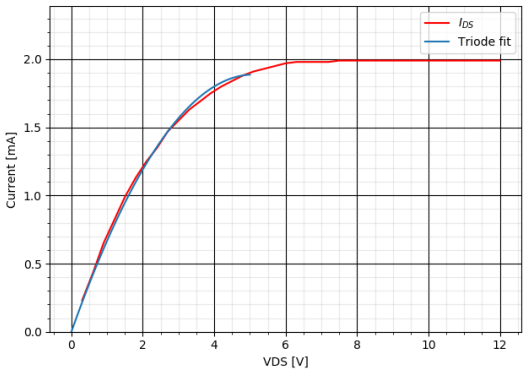

Figure 3.3: I-V fit in triode region

Iterating the procedure for the other devices, the following μC_{ox} and R_{lin} values are obtained:

The analysis of the data confirms that the value of $\mu \cdot C_{ox}$ is independent of the size of the transistor, while the drift region influences it. The values of

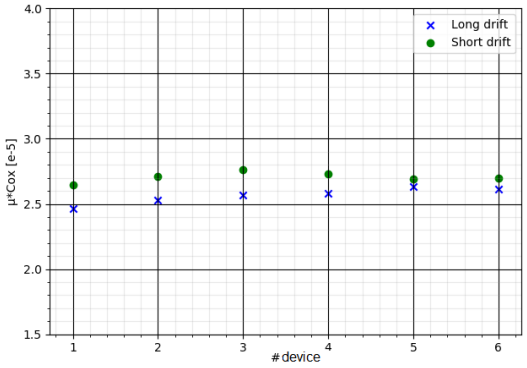

Figure 3.4: Extrapolated $\mu \cdot C_{ox}$ in triode region

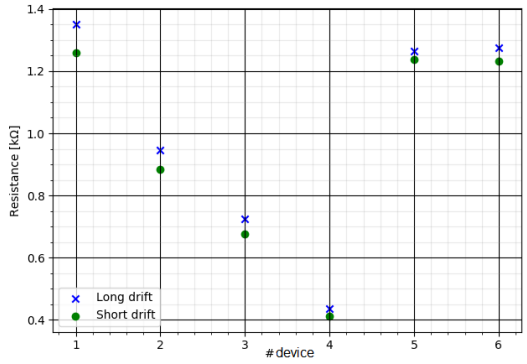

Figure 3.5: Extrapolated linear resistances

Device	long channel	short channel
1	1350	1259
2	945	882
3	726	675
4	436	412
5	1265	1237
6	1275	1233

Table 3.2: Linear resistance values for each device (in Ω)

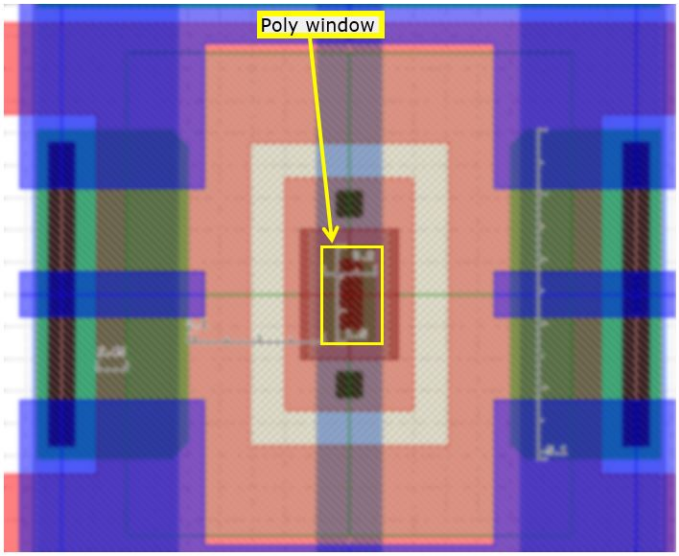

Figure 3.6: Bit-cell Layout

$\mu \cdot C_{ox}$ are then used to extrapolate the linear resistances, the value of which is collected in the table 3.2, according to the equation:

$$R_{DS} = \frac{1}{\mu C_{ox}} \cdot \frac{L}{W} \cdot \frac{1}{(V_{gs} - V_{th})} \tag{3.2}$$

The drift region also affects the resistance values: a longer drift region makes the device more resistive. No wonder: the drift transistor is specially used to withstand high voltage stresses. The drift region is in fact an additional restitive path to limit the effective electric field.

3.2 Dielectric

Now the focus is shifted to the next row of wafer pads to test the real bit-cells: the previous transistors plus the oxides, as in Fig. 3.2.

Although the anti-fuse OTP will be programmed under constant voltage stress, a DC voltage ramp could be a useful experimental method for collecting wear current data over a wide voltage range. First the access transistor is turned on. Then, an increasing DC ramp from 0 to 10 V is applied to the anti-fuse element. The current flowing in the bit-line is measured for

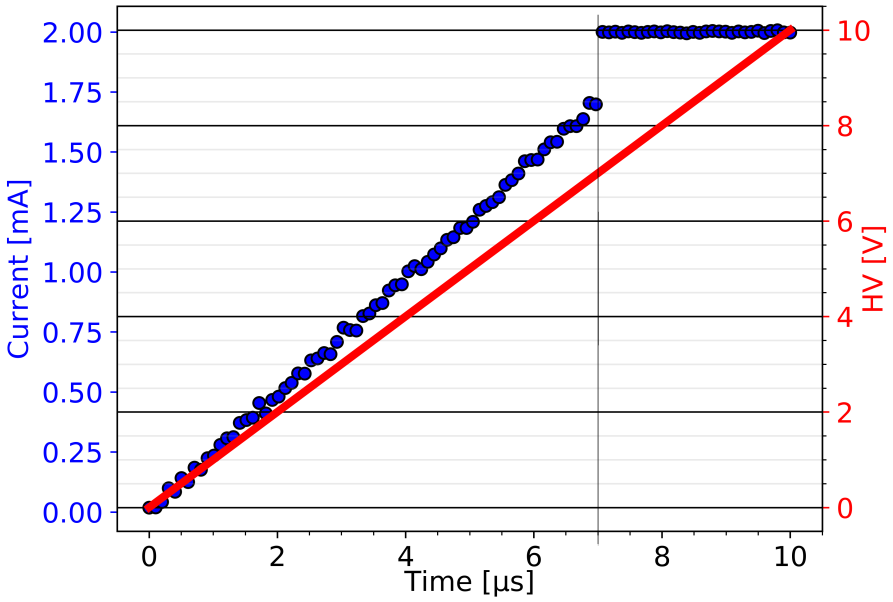

Figure 3.7: DC voltage ramp test (data missing nearby T_{BD})

each voltage value. The wearout current is, in fact, the leakage current of the stressed dielectric. When HV is above a certain critical value, the break occurs within $\simeq nS$. Therefore, it would be difficult to gather current data near that point. To some extent, the cumulative feature of the BD event could simplify the situation. In fact, a dielectric subjected to a voltage stress of duration T, reacts identically to how it would react in case of multiple shorter stresses, T_i, so that $\sum_i T_i = T$.

Another technique is that of Fig. 3.1 (right). Short high voltage pulses are applied separately to each bit-cell. Leakage currents are persisting up to the breakdown of the capacitor. The same tests are performed using arrays with many identical cells to have a more meaningful statistic. Furthermore, the procedure allows to calculate the T_{BD} for each dielectric value and HV. Unexpectedly, when dealing with ultra-thin oxides, the wearout current for some HV values was equal to the saturation current ST, an additional limiting factor of current data collection.

The wearout current theory discussed in Chapter 2 already suggests that FN tunneling should be dominant for the 7.7 nm oxide. For 2.2nm oxide things get more complicated, but direct tunneling should still contribute

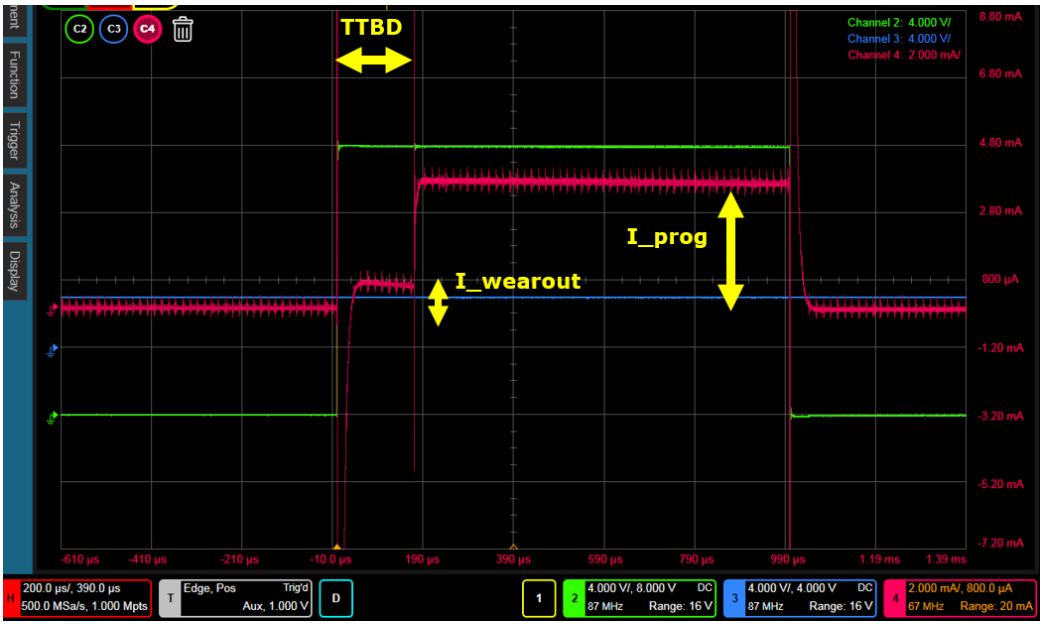

Figure 3.8: Example of BL current transient and parameters

more. The tunneling current equations ensure that the magnitude is inversely proportional to the thickness of the oxide. Reasonable: in fact the thinner the oxide, the greater the vertical electric field with the same HV value. The greater the electric field, the greater the likelihood of tunneling. The tunneling current should also be proportional to the surface of the oxide ($W_{ox} \cdot L_{ox}$) for the same reason: the larger the surface, the greater the electric field.

Referring to the devices studied in Table 3.1, it is therefore expected that the tunneling grows from device #1 to #4 and then again, given the increase of L_{ox} , for #5 and #6 devices.

Both wearout and T_{BD} data were collected by applying a 3.3 V pulse to the ST gate and biasing the BL at [13 , 16] V for the 7.7 nm oxide and [7.8 , 5] V for the 2.2 nm one. The oscilloscope screenshot of Fig. 3.8 provides a representation of how to collect the desired data.

The system should be able to clearly detect a current of the order of $10\mu A$, with a sampling rate of at least $\simeq$ MHz. These conditions are quite challenging: a more convenient solution for faster data collection such as the use of an automated probe station should be discarded as it is not accurate

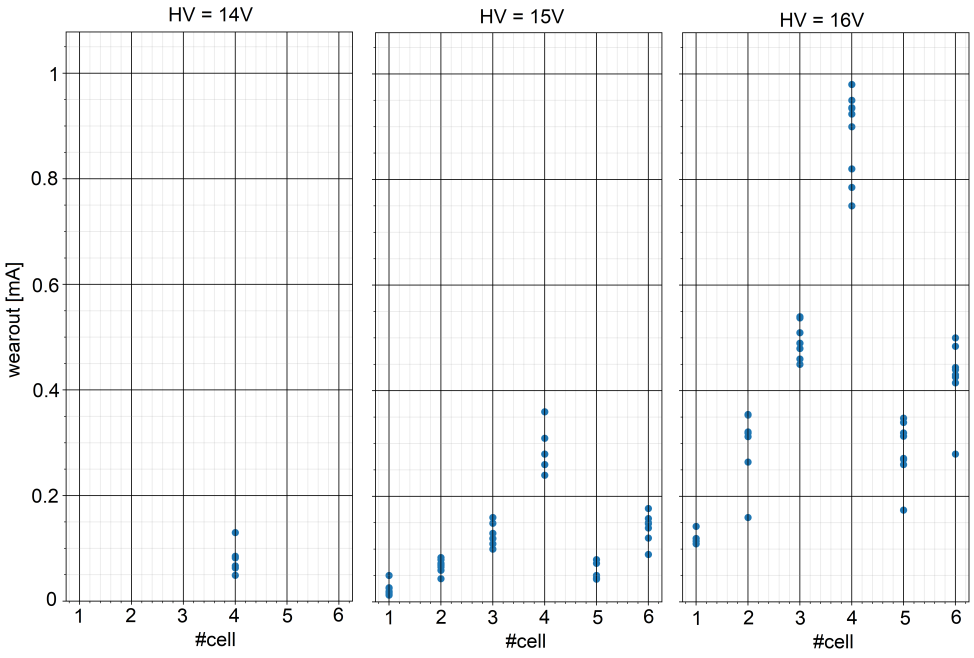

Figure 3.9: Wearout data collection

enough. In fact, that is more often used for evaluating the lifetime of the device (sampling time of milliseconds, not nanoseconds for sure). For this reason, a manual pico-probe station with a sensitivity of μs was used, despite the time required for measurement and a higher probability or human error. Hundreds of cells have been tested. The data population may not seem entirely satisfactory, however the results were promising and appeared to be in agreement with physical models of wearout and T_{BD}.

7.7nm oxide

The wearout data collected by varying the HV for each cell are shown in Fig. 3.9. For HV = 14V, measurements had such high noise that data collection was only reliable for the largest cell. Fig. 3.10 shows the average wearout values for HV=15 and 16V. In fact, data was also collected for 13 and 14V, but the corresponding current was too small ($\simeq 1 - 30\mu A$)to be precisely measured by the presented setup, where the noise (long cables etc.) superimposes white oscillations μA on the real value.

Always referring to Table 3.1, both W_{otp} and W_{ST} change from cell #1

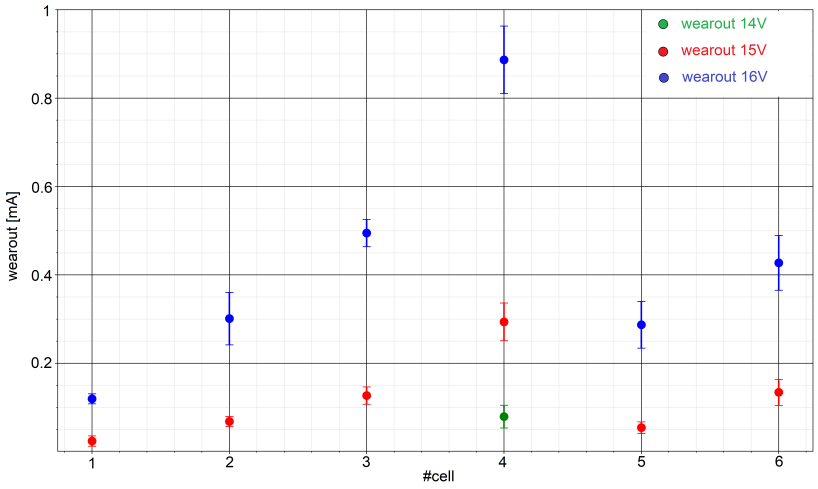

Figure 3.10: Average wearout vs #cell

to #4. Thus, cells #5 and #6 differ from #1 for L_{otp}. In fact, the difference in ST dimensions is not expected to have an appreciable effect on the wearout current. To empathize the impact of the dimensions, the data were reordered in Fig. 3.10 by oxide area. The data already reveal little evidence: the wearout current is proportional to the OTP area (both ST and oxide). Furthermore, it appears to be directly proportional to the applied voltage. The first statement deserves some discussion of a few words. The tunneling current is quite low: the ST works in its linear region. The resistance, R_{lin} is known from preliminary ST experiments. At the same time, the oxide also acts as a resistor, $R_{wearout}$, certainly larger than the ST one (otherwise the current would have been $\simeq$ that of saturation). In this way the cell can be thought of as composed of two resistors connected in series as in Fig. 3.12. R_{lin} and $R_{wearout}$ respectively depend on the size of the selector transistor and of the oxide. If the first dependency: $R_{lin} \propto L_{ST}/W_{ST} = 0.45\ mum/W_{ST}$ is quite clear (eq. 3.2), the second is more complicated. Two hypotheses:

$$\begin{aligned} &- R_{wearout} = R_{lin} \cdot 100 \\ &- R_{wearout} \propto 1/A_{otp} = 1/\big(W_{otp} \cdot L_{otp}\big) \end{aligned} \tag{3.3}$$

Based on this assumption, the total bit-cell resistance for cell #1, R_{tot_1}, can be estimated as follows:

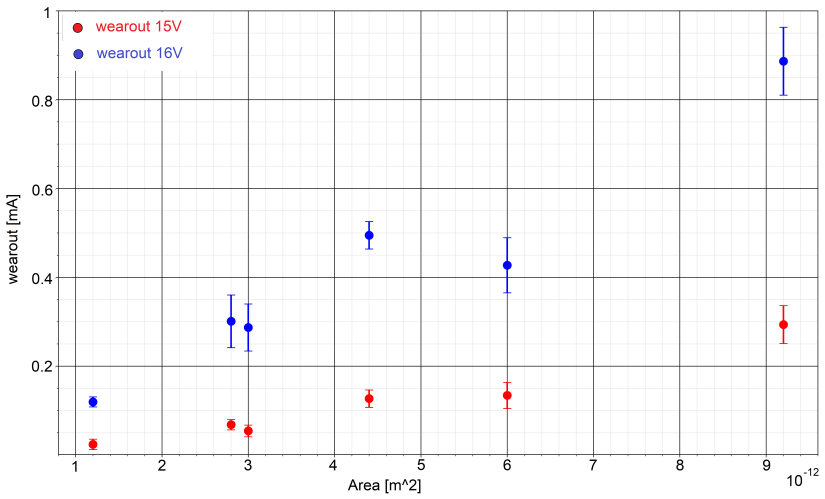

Figure 3.11: Average wearout vs cell area

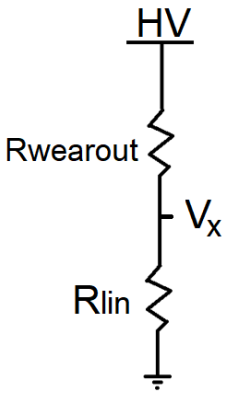

Figure 3.12: Bit-cell model during wearout phase

$$R_{tot_1} = R_{wearout_1} + R_{lin_1} = R_{wearout_1}\Big(1 + 1/100\Big) \tag{3.4}$$

Therefore:

$$I_1 = HV/R_{tot_1} \propto HV/R_{wearout} = 16/(100 \cdot R_{lin_1}) \tag{3.5}$$

From Fig. 3.10, $I_1 \simeq 0.12mA$, which means $R_{otp} = 133k\Omega$, proving the validity of the third hypothesis: in fact $R_{lin_1} \simeq 1.3$ in Fig. 3.5.

As a further proof, in support of the above results, the current I_2 can be predicted by exploiting the well-known ratio between the cell areas #1 and #2 OTP:

$$R_{wearout_2} = R_{wearout_1}/\big(W_{otp2}/W_{otp1}\big) \tag{3.6}$$

as well as:

$$R_{lin_2} = R_{lin_1}/\big(W_{ST2}/W_{ST1}\big) \tag{3.7}$$

Therefore, substituting the values from Table 3.1:

$$R_{tot_2} = R_{wearout_2} + R_{ST_2} = \frac{R_{wearout_1}}{\big(3,5\mu m/1,5\mu m\big)} + \frac{R_{ST_1}}{\big(7,1\mu m/5,1\mu m\big)} \tag{3.8}$$

$$R_{tot_2} \propto R_{wearout_1} \cdot 0,43 \tag{3.9}$$

This procedure leads to:

$$I_2 = HV/\big(R_{wearout_1} \cdot 0,43\big) = 16/\big(43\cdot)R_{ST_1}\big) \approx 0,286mA \tag{3.10}$$

Checking Fig. 3.10, this value is within one standard deviation. The same procedure can be iterated on the other devices, with the general formula:

$$R_{tot_n} = R_{wearout_n} + R_{lin_n} = R_{wearout_1}/x + R_{lin_1}/y = R_{wearout_1}\Big(1/x + 1/100y\Big) \tag{3.11}$$

Where x and y are the increments of A_{ox} and W_{ST} with respect to cell #1 (since the L_{ST} technology is constant):

$$\begin{aligned} x &= \left(\frac{W_{otp_n} \cdot L_{otp_n}}{W_{otp_1} \cdot L_{otp_1}}\right) \\ y &= \left(\frac{W_{ST_n}}{W_{ST_1}}\right) \end{aligned} \tag{3.12}$$

For example, knowing I_1, the current I_5 of device #5 should be 2.5 times larger since the only change would be in the L_{otp} parameter, from $0.8\mu m$ to $2\mu m$. Proof:

$$\begin{aligned} &R_{tot_5} = R_{wearout_1}\Big(0,8/2 + 1/100\Big) = R_{wearout_1} \cdot 0,4 \\ &=> I_5 = I_1/0.4 = 0,32mA \end{aligned} \tag{3.13}$$

Always in accordance with the average data collected. The iteration on all cells (even for HV = 15V) is consistent with the collected data, demonstrating the functionality of the cell model and successfully concluding the discussion on the direct proportionality between the wearout current and the OTP area.

Wearout modeling

The next step would be to demonstrate the validity of the Fowler-Nordheim tunneling current model (the electric field is $> 15\,\mathrm{MV/cm}$, the FN current dominates). To fit the data, an exponential formula based on FN theory is used:

$$I_{FN} = A \cdot V^2 \cdot exp\Big(\frac{-B}{V_{cap}}\Big) \tag{3.14}$$

However, the population of the data is far from significant: the setting was limited to two points per cell, with the exception of cell #4. At least 3 points are required to obtain reliable A and B parameters or any curve would fit the data. Therefore the technique is to extrapolate those parameters from cell #4 and use them to also fit the data of other cells as proof of their correctness.

The extrapolated A and B parameters are:

$$A = 3.5 \cdot 10^{-3}\,\mathrm{A/V^2} \text{ and } \;\; B = 3.8 \cdot 10^{10}\,\mathrm{V/m} \tag{3.15}$$

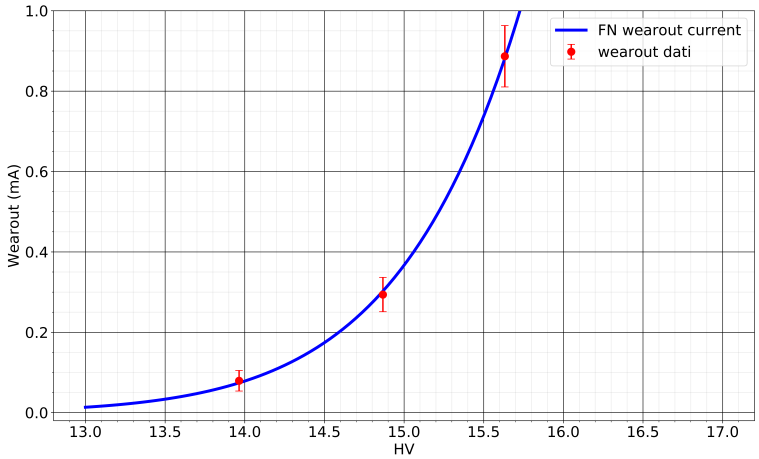

Figure 3.13: FN fit for cell #4

The results obtained could also be used to predict the oxide wearout current of the 2.2nm oxide. In fact, the two oxides are almost identical, with the only exception being their thickness. Therefore parameters A and B should still be valid. However, the resulting data extrapolated with this technique may differ from the actual wear current of 2.2 nm. The reason is that the dominant tunneling mechanism may be the direct one this time, rather than the FN, having two slightly different equations. More generally, for 2.2 nm oxide wearout data, $I_{wearout_{tot}} = I_{direct} + I_{FN}$ where $I_{direct} >> I_{FN}$. In fact, the threshold between FN and direct tunneling depends on the electric field. Considering a threshold of E = 15MV / cm:

$$E = 15MV/cm = \Delta V/Tox => \Delta V = 1.5e9 \cdot 2.2e - 9V = 3.3V \quad (3.16)$$

This means that above an HV stress of 3.3 V the FN becomes the dominant mechanism. Therefore the following prediction should not be too far from the measurements.

To correct the idea, the above experiment is calculated on the #1 cell, with a voltage drop of 5V.

$$I_{FN} = W_{otp_1} \cdot L_{otp_1} \cdot A \cdot (5V/2.2nm)^2 \cdot exp(-B \cdot 2.2nm/5V) = 1.32\text{A} \quad (3.17)$$

The sought points are then calculated by Phyton script by varying the HV. The result is represented in Fig 3.14.

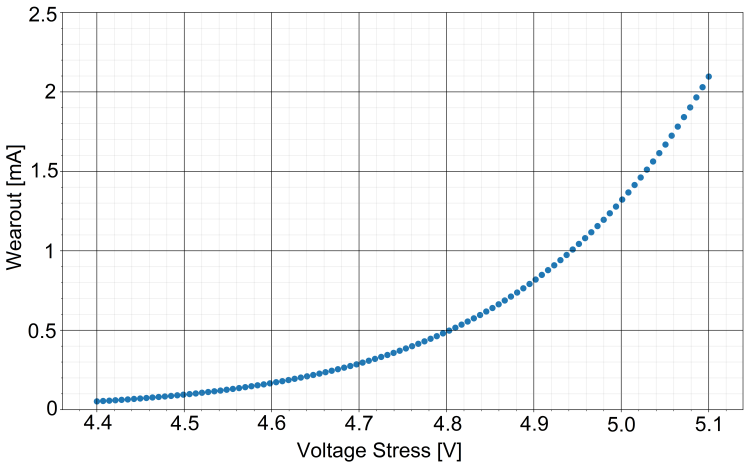

Figure 3.14: Wearout current prediction for #1 of 2.2 nm oxide using FN results in 7.7 nm

In carrying out the above calculation, to be precise, instead of using the high voltage applied to the cell as abscissa, the effective oxide voltage was used, subtracting the voltage drop of ST. It is interesting to note that for each wear value corresponding to a stress higher than 5.1 V, the tunneling current FN already reaches the ST saturation current (Fig. 3.3). Therefore, the tunneling current is expected to be so high that ST works in the saturation region already at low voltage.

2.2nm oxide

The measurements related to 2.2nm oxide are now analyzed, with the same approach used previously for 7.7nm. The first data are collected (Fig. 3.15) for each cell by varying the HV from 7.5 to 9V. However, it is important to point out that the actual voltage stress across the oxide terminal could actually be much lower, possibly between 3.5 and 7V depending on the ST size. Indeed the wearout current is now much higher than before and the 3.2 hypothesis previously used in the 7.7 nm oxide calculations is no longer valid:

$$R_{wearout} \neq 100 \cdot R_{ST} \tag{3.18}$$

Indeed, the equivalent resistance of the OTP cell should now be compara-

ble to R_{lin}. The voltage drop of the oxide would therefore be obtainable with the law of the resistive divider. The hypothesis then becomes $R_{wearout} = R_{lin}$, while 3.2 reduces to:

$$R_{tot_n} = 2 \cdot R_{lin_n} = R_{lin_1} \cdot \Big(1/x + 1/y\Big) \tag{3.19}$$

The above formula allows to predict the wearout current for all the other cells starting from the wearout current of the #1 cell (as done for the 7.7 nm oxide).

For example, the current I_2 when HV = 7.5V is calculated as follows (note: $L_{otp_1} = L_{otp_2}$):

$$\begin{aligned}
&I_1 = 1,5mA = HV/R_{tot_1} = HV/(2 \cdot R_{ST_1}) \\
&1/x = 1,5/3,5 \\
&1/y = 5,1/7,1 \\
&=> R_{tot_2} = R_{otp_2} + R_{ST_2} = R_{otp_1} \cdot \left(\frac{1,5}{3,5}\right) + R_{ST_1} \cdot \left(\frac{5,1}{7,1}\right) \\
&\text{fromHP :} \\
&R_{tot_2} = R_{ST_1} \cdot 1,15 \\
&=> I_2 = I_1 \cdot 2/1,15 = 2,61
\end{aligned} \tag{3.20}$$

While for the cell #5 it would be $I_5 = 2,4mA$. In reality these results are not exactly compatible with those obtained through experiments (especially the last one: the #5 cell should have a maximum current around 1.7mA). The error lies in the assumption that the ST still behaves as a resistance R_{lin}, while instead it works at least in the quadratic region. So each HV step actually leads to an increase in the ST resistance. This makes $V_x = I \cdot R_{ST}$ no longer valid.

To have a better understanding and correlate the data, the measurements were averaged and plotted in Figure 3.16.

The current flowing in the device #1,#5 and #6 is comparable (within one standard deviation). This is different from what was achieved for the 7.7nm oxide. An increase in the oxide area should result in a lower equivalent resistance and therefore a greater $V_x = V_{ds}$. In other words: more current. The ST, considering the voltage drop of the oxide, is working on the quadratic region, a little before the saturation threshold, giving some leeway to further

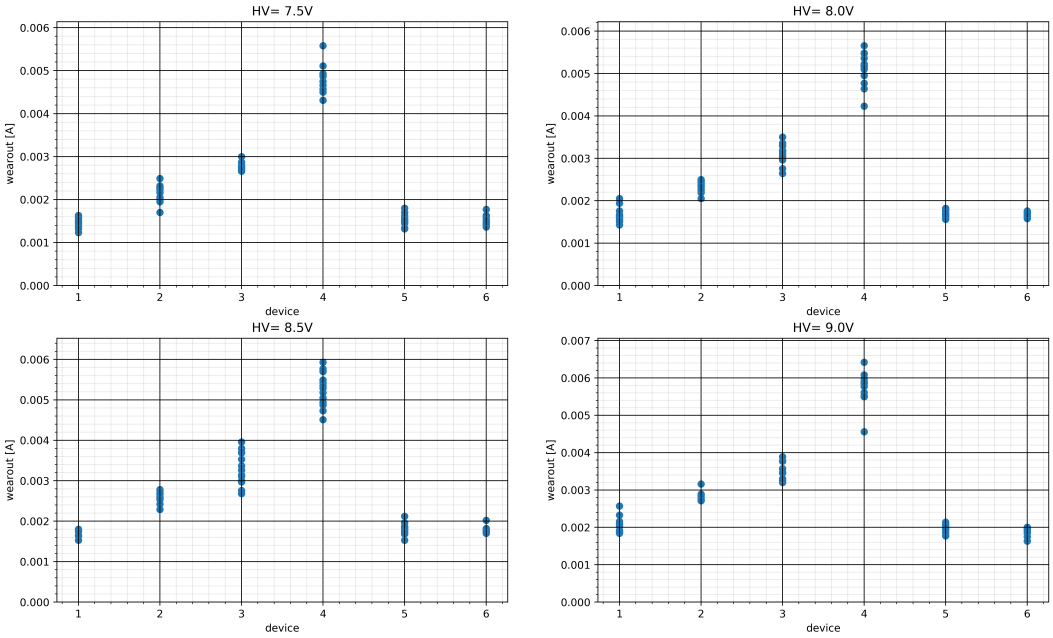

Figure 3.15: Collection of current wear data for each device varying HV

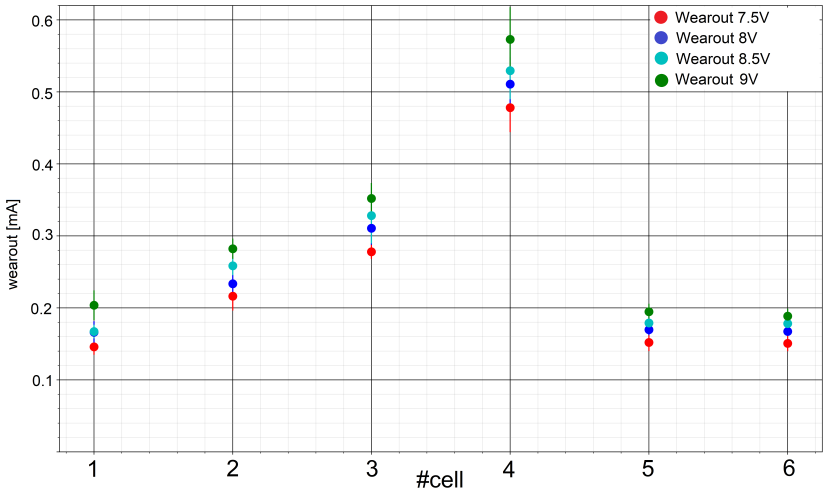

Figure 3.16: Average wearout current for each device varying HV

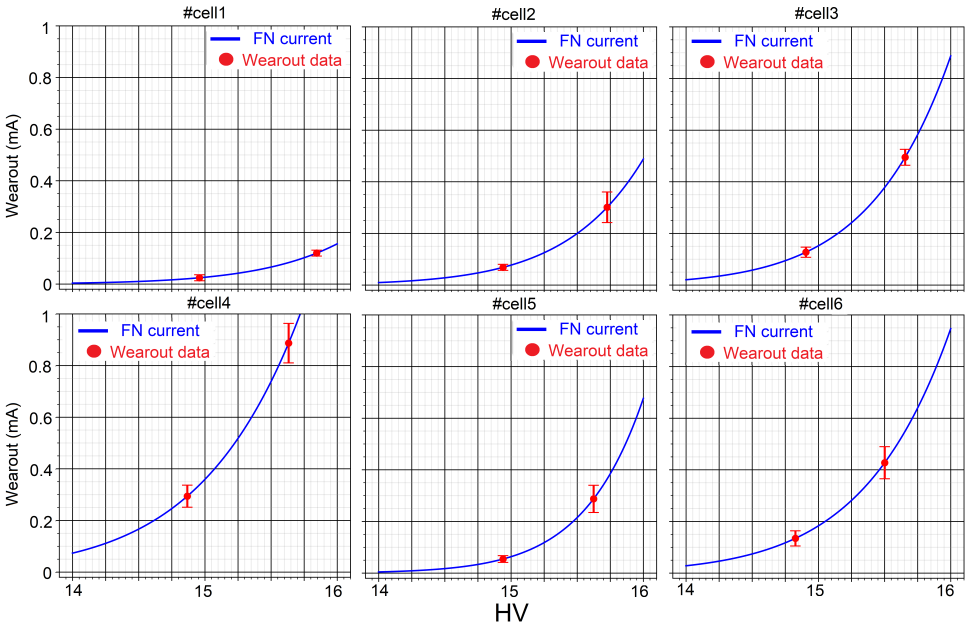

Figure 3.17: Wearout data fit for each cell

increase the current. With a more consistent statistic this effect could have been identified.

Regarding the 7.7 nm, the data are fitted with the wear current model FN. The abscissa refers to the actual extrapolated oxide voltage as if there was virtually no ST.

Although the fit seems comforting, the extrapolated parameters are not very significant. In fact in most cases the current saturates due to the presence of ST, therefore the wearout data could be greater. In fact, the parameters A and B do not even correspond to those obtained from the analysis of the 7.7nm oxide.

Another detail in 3.16 is that the #5 and #6 devices give the same results, making the wearout current independent of L_{otp}, contrary to the FN prediction. Once again this is related to the saturation limit imposed by the ST. In fact, taking the FN parameters extrapolated from cell #5 as valid, other cell currents can be foreseen. Fig. 3.18 shows the result of this operation, which actually support the hypothesis that cell #5 and #6 should have different values, while the correct values are guessed for the other cells.

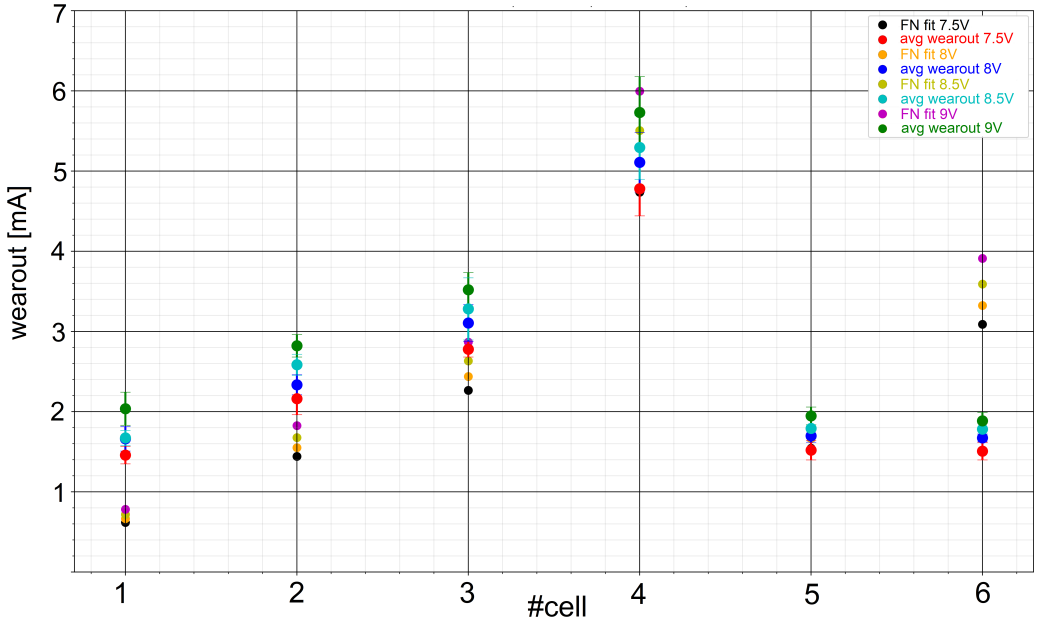

Figure 3.18: Using the parameters A, B from cell #5 to predict the current of the other device

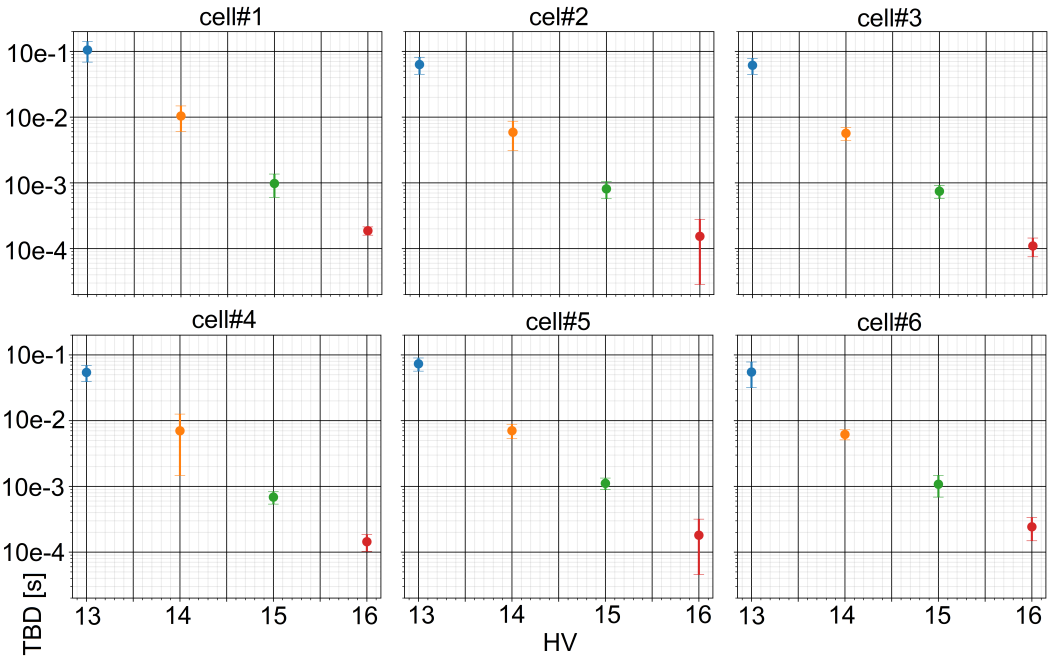

Figure 3.19: Average of T_{BD} measurements vs # cell

3.3 Time To Breakdown T_{BD}

Through a Phyton script, the current transients just exploited for the wearout current analysis are also used to collect the T_{BD} data. The goal is to validate the models that predict T_{BD} as a function of oxide thickness and applied voltage. Three models to test: the power law, the E model and the 1/E model. T_{BD} is expected to depend on: the HV pulse, the oxide size W_{otp}, L_{otp} and the thickness. Basically the same dependencies as the wearout current, although the curve shapes are different. Since T_{BD} has nothing to do with ST, its size, technology or drift region, the data collected for the length of the different drift region is averaged together to yield higher samples.

3.3.1 T_{BD} 7.7nm

The average T_{BD} for each cell varying the HV is shown in Fig. 3.19.

Some evidence: as the applied voltage increases, the T_{BD} necessarily decreases. The same happens when the oxide area is increased. Moreover, from the semi-logarithmic graph of Fig. 3.19, there seems to be an exponential

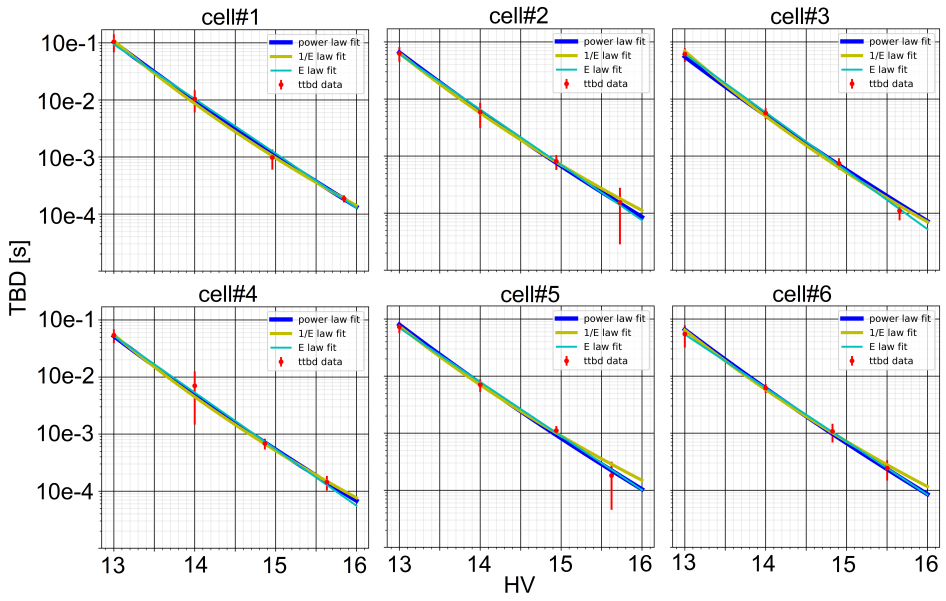

Figure 3.20: T_{BD} models fitting

relationship between T_{BD} and HV. The above data are fitted using the following equations:

Power law model: $A \cdot (V_{cap})^{-B}$

1/E model: $const \cdot exp(G \cdot T_{ox}/V_{otp} \cdot exp(E_a/(K_b \cdot T)) = C \cdot exp(D/V_{cap})$

E model: $const \cdot exp(-G \cdot V_{otp}/T_{ox}) \cdot exp(E_a/(K_b \cdot T)) = E \cdot exp(-F \cdot V_{cap})$

Fig. 3.20 shows how well the equations fit the data for each cell.

The extrapolated fitting parameters, returned by the script, are those in Fig. 3.21

The analysis did not lead to the expected results: the arguments of the exponential / potential terms B, D, F should be mathematically independent of the OTP area. In fact, the dependence should already be taken into consideration in the proportional terms A, C, E. However, the adaptation suggests the opposite: the parameters change from one device to another.

Since the error is due to the algorithm used, another strategy is required.

```
POWER LAW:
[[ 4.25e+34   -3.19e+01]
 [ 5.70e+33   -3.13e+01]
 [ 1.91e+36   -3.36e+01]
 [ 5.55e+34   -3.23e+01]
 [ 1.14e+33   -3.07e+01]
 [ 1.29e+33   -3.08e+01]]

1/E MODEL:
[[4.27e-17    4.61e+02]
 [1.24e-16    4.40e+02]
 [7.48e-18    4.77e+02]
 [3.07e-17    4.56e+02]
 [2.92e-16    4.31e+02]
 [1.48e-16    4.38e+02]]

E MODEL:
[[2.45e+11    2.19e+00]
 [2.53e+11    2.23e+00]
 [1.25e+12    2.35e+00]
 [4.18e+11    2.28e+00]
 [1.42e+11    2.18e+00]
 [8.07e+10    2.15e+00]]
```

Figure 3.21: Extrapolated T_{BD} fit parameters

A good idea would be to solve the problem as follows: parameters B, D, F will henceforth be kept fixed for all cells and set to be equal to the statistically most significant fit. Thus, the only dependence on the area is forced to be on the terms A, C and E. The differences are evident, especially for the 1 / E model. The graph and the updated parameters are shown in Fig. 3.23.

Generally, the literature suggests a $B = 43 \pm 2$, while here it is set to 32. This would lead to a mismatch of at least 20% and some discrepancies with other works. However, as far as the E model is concerned, the results match up quite well.

However, the experiment is still not sufficient to determine which of these defect generation mechanisms has the greatest impact. In fact, breakdown is generally a consequence of multiple overlapping mechanisms. However, predictions on T_{BD} can still be obtained by extrapolation.

3.3.2 T_{BD} 2.2 nm

Probably due to the tendency of the ultra-thin oxide to settle down on a local minimum when the oxide is only partially degraded (also called soft-breakdown), the T_{BD}, contrary to what happened in the case of 7.7nm, seems avoid any addiction. The results of T_{BD} range from μs to tens of ms. Some-

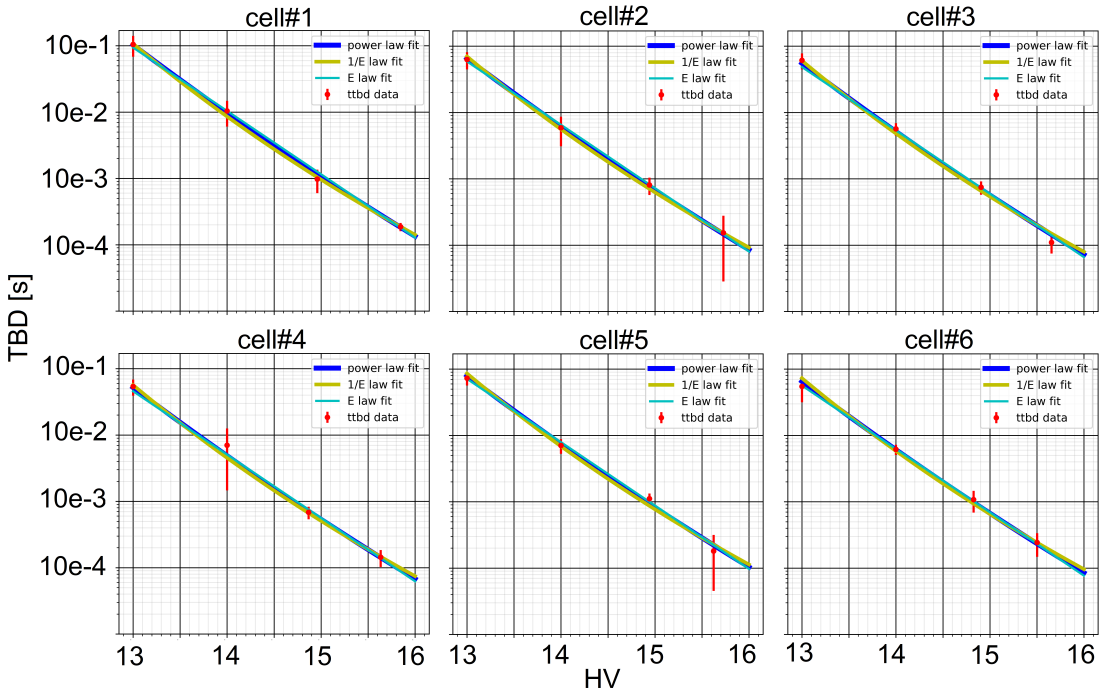

Figure 3.22: T_{BD} fit with constant B, D, F parameters (extrapolated from device #1)

```
POWER LAW:
[[4.56e+34    3.20e+01]
 [2.89e+34    3.20e+01]
 [2.44e+34    3.20e+01]
 [2.31e+34    3.20e+01]
 [3.53e+34    3.20e+01]
 [2.89e+34    3.20e+01]]

1/E MODEL:
[[4.61e-17    4.60e+02]
 [3.00e-17    4.60e+02]
 [2.59e-17    4.60e+02]
 [2.43e-17    4.60e+02]
 [3.69e-17    4.60e+02]
 [3.13e-17    4.60e+02]]

E MODEL:
[[2.47e+11    2.20e+00]
 [1.55e+11    2.20e+00]
 [1.27e+11    2.20e+00]
 [1.20e+11    2.20e+00]
 [1.88e+11    2.20e+00]
 [1.49e+11    2.20e+00]]
```

Figure 3.23: Extrapolated fit parameters with constant B, D, F

times, even comparing the measurements for different HV stresses, the results can be similar or even returning higher T_{BD} for higher HV. The hypothesis is that the soft breakdown can be considered as a local equilibrium point where the system is stable for a certain period of time. During the soft-break the ohmic path is not yet completed and the quantum tunneling is still contributing to the overall current. It can be thought of as oxide damage (which is only partially broken) whose effect is to decrease the equivalent resistance. In the meantime, other soft failures can be created, resulting in more resistors in parallel. However, once the oxide eventually breaks down, the current will be purely ohmic and electrons will only flow through the latter lower ohmic path.

The 2.2 nm oxide also exhibits an invasive bulk current flowing that anticipates the rupture event. Many experiments in the literature focus on this bulk effect. While this effect is secondary, it must be taken into account when designing the OTP chip to avoid over-currents. From the physical point of view, the phenomenon could be explained as follows: once the HV pulse is applied, due to the effect of the electronic tunneling through the oxide, a thermal effect is activated which generates holes inside the dielectric. These holes are directed towards the bulk generating a positive step in the current

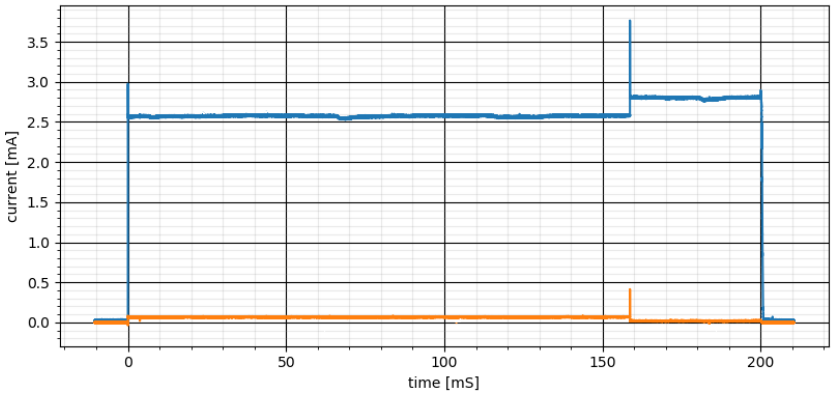

Figure 3.24: BL (blue) and bulk (orange) current transient

BL.

To proceed with the experiments, the cell was tested by separating the bulk and source terminals. In this way the breakdown event is more detectable and T_{BD} is measured when the ground current drops to 0, which means that there is no more tunneling current. The figure 3.24 shows the separate bulk and source currents. The breakdown occurs around 160 ms.

3.4 Evaluation of R_{OTP} after programming

The experiments have been concluded. The order of magnitude of the wearout current and T_{BD} were found. What is still missing, in order to have all the important specifications for the circuit design, is the resistance value of the rupture oxide. In fact, the reading circuit (sense amplifier) must be able to detect which bit-cell is programmed and which is not. The distinction is based on the expected current flowing for a given voltage applied to the BL. Therefore, to fix the reference threshold, it would also be necessary to have an order of magnitude of post-breakdown oxide resistance. The creation of a chain of defects that characterizes the breakdown of the anti-fuse is, for an ideal Brownian oxide, very close to a random walk. The first defects are generated inside the oxide with a certain speed but randomly in space. However, when the oxide is weak and damaged, a local area of higher electric field will be created, making weak point breakage preferable. In general, there are many ways to form a defect chain: it can be straight thicker etc. as explained

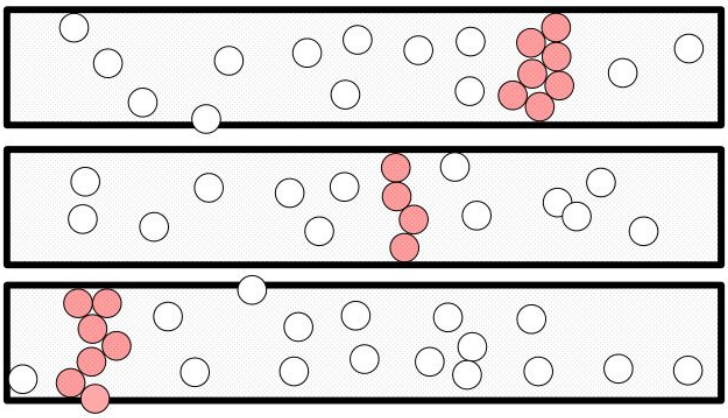

Figure 3.25: Dielectric breakdown paths

in [9].

Statistically, there will be an average R_{OTP}, coinciding with the most likely way to create the conductive path, and a large standard deviation representing the many other possible ways to create the low-ohmic path. The resistance distribution would be a very smooth Gaussian.

The resistance should be dependent on the thickness of the oxide and greater for the 7.7 nm oxide: in fact the defect chain will certainly be longer and therefore more ohmic. Also, the HV stress used during programming should not affect the value. In fact, the greater the electric field, the greater the generation of defects per unit of time, but without any relationship with their spatial density.

3.4.1 Results

The calculation of R_{OTP} was carried out both for the thickness of the oxide: 7.7 nm and for 2.2 nm. Since the drain node ST is not directly accessible, the values are extrapolated, using once again the precious data R_{lin} of Fig. 3.5. First, the actual data for a HV = 1V are collected (simulating a reading phase) for all the cells, previously programmed in different conditions. Hence, the total resistance of the OTP cell is obtained by making the reciprocal $(1V/I)$. Subtracting the R_{lin} corresponding to each device gives R_{OTP}. The average of the device size and the length of the drift region gives: $3.5k\Omega$ for the 7.7 nm oxide and $2.7k\Omega$ for the 2.2 nm oxide.

Values as expected showed no area correlation. Furthermore, the programming condition (such as the HV stress used) did not affect the ohmic spot. However, it must be understood that the number of samples collected is a few hundred measurements per oxide which may not be sufficient to have a consistent average value, but certainly high enough to have the order of magnitude: $\simeq k\Omega$. Furthermore, a more appreciable impact of the oxide thickness on the R_{OTP} value was expected: for a factor of 3.5 in the thickness there is only a 30% difference in the resistance value. This could be related to the very wide distribution of data found: individual data was spreading widely, making it difficult to obtain data consistency.

Finally, the $\simeq \mu A$ offset setting was making everything more challenging and worsening the already delicate situation.

Chapter 4

OTP Memory Chip Design

4.1 Introduction

Reliable OTP memory for the trimming application requires a very low error rate. Of course, an error detection algorithm will be implemented to further minimize the failure of the chip to reach the typical value of 0.01 ppm. The most likely failures derives from poor oxide programming. The anti-fuse, in fact, must be hardly broken down, to avoid misunderstandings when reading the memory. A threshold, as anticipated in Chapter 3, derives from the average value of R_{OTP} ($3.5k\Omega$ for example). The sense amplifier, during the reading phase, will force a defined current, $1\mu A$, through the cell. The voltage across the oxide is then read, if the value largely exceeds $1\mu A \cdot 3.5k\Omega$ it means that the oxide for some reason was not hardly broken, but only damaged and the reading circuit could confuse the bit as not programmed. The greater the programming electric field, the greater the likelihood of having hard breakdown in the programming time window, $100\mu s$, in this application. However, producing a higher voltage would require more power and area consumption, affecting the overall cost of the chip. An oxide of 7.7 nm, to meet the failure rate specification, may actually need 18V: the car battery would not be enough and a bulky charge pump should therefore be added to the design. This is why, despite the fact that Chapter 3 returned less consistent results, it is worth continuing the design of the OTP memory using 2.2nm oxide, with which the voltage of the car battery can guarantee a reliable and quick programming phase. In particular, the #2 cell is considered to be a good candidate to become the destination OTP bit cell.

Oxide	Programming Voltage	Programming window	Failure rate
2.2nm, cell #2	<13.5V	100μ s	0.01ppm

Table 4.1: OTP Memory specifications

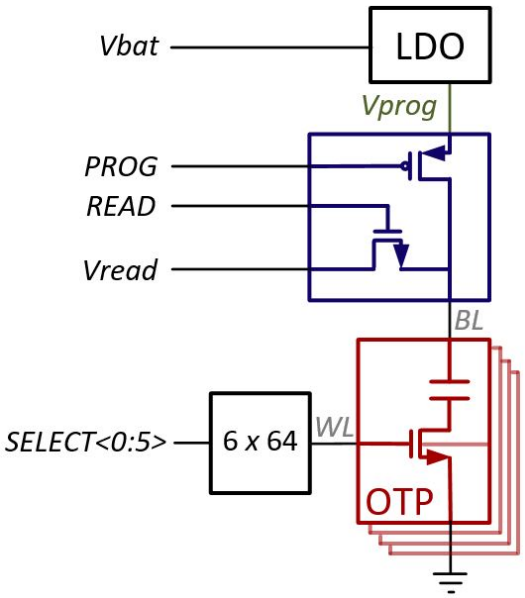

Figure 4.1: Top level OTP memory concept

Once the choice of the anti-fuse element has been thoroughly studied, the design of the circuit begins. The basic circuit concept of the higher level OTP memory is shown in Fig. 4.1. It consists of: an OTP module, a programming voltage regulator and a reading circuit. A decoder will be used to select the Wordline to program / read.

The purpose of the concept above is to have a first look at the behavior of the cell with an appropriate driver circuit, no longer ideally or externally supplied. During the programming phase, the vehicle battery voltage is supplied to the LDO regulator which produces the programming voltage of $10.7V$. A $PROG$ signal activates programming via a switch, which essentially creates the connection between the BL cell and the LDO regulator. The 6-bit decoder is used to select the cells of the desired WL once at a time. The reading phases, on the other hand, require the $READ$ signal to change the mode and an externally supplied voltage V_{read}, applied to the BL to read the contents of the cell. The $Vread$ voltage is secondary, but critical if not specified correctly. In fact, $Vread$ depends on the oxide quality experiments: the higher the voltage, the easier it is to detect the flowing current and the easier it would be to design the sense amplifier (requiring less sensitivity). However, a high $Vread$ could lead to involuntary cell programming after many reading phases: this must definitely be avoided. $Vread < 2V$ is an additional specification beyond those in Table 4.1. While the chip is in the layout and production stage, more sophisticated features are being studied and designed for the second and more comprehensive test-chip, which is expected to be ready in March 2022. Similarly to the final product, it will also contain the sense amplifier for reading, an error detection algorithm, an internal reference voltage generation (bandgap generated [27], [28]), observers for line open-circuit and short-circuit detection, iDAC for failure rate control and many other secondary circuits.

This chapter presents the top-level circuit, sketched in Fig. 4.1. The circuit is designed, taped-out and the test-chip tested by the author with a 20-pin ceramic package. The core of the chip, which required a lot of attention, is the LDO regulator.

4.2 LDO Regulator

This section presents a dual domain capacitor-less Low-DropOut (LDO) voltage regulator developed to drive 512 OTP memory cells. The LDO is powered by the car battery voltage and generates a constant OTP programming voltage, regardless of variations on the line and equivalent load capacity and resistance. The LDO consistently provides 10.8 V output with a nominal error of 0.5% while consuming $500\mu A$ idle current and is capable of delivering up to 10 mA at a variable resistance load of $10pF$. The settling time is $< 1\mu$ s at transient full load without presenting overshoot and ringing which would cause a more rapid and uncontrolled deterioration of the cells. An average current of $2.5mA$ flows in a programmed cell during the programming phase, while $60\mu A$ flows during the reading phase, making it distinguishable from a virgin cell. In a 0.35μ m CMOS technology, the LDO occupies an active area of 220x324 μ m ^{2}. The measurements validate the proposed structure.

The OTP module consists of a WordLine X BitLine matrix (64x4) with a bit cell in each node. A decoder is used to point to a specific WordLine for reading or programming.

The LDO works between two domains. $V_{dd} = 3$V to constantly bias the low side Error Amplifier (EA) and $V_{bat} = 13.5V$, the voltage of the battery, used to power the output stage, composed of a PMOS switching device and a resistive feedback. In fact, V_{bat} varies a lot, as in the cold start of the car, making it inappropriate to bias the EA as in [4]. In fact, a 3V domain is also useful for generating low-side and digital circuits for the following applications. Also, having a low-side EA helps to minimize area occupation and matches the current low-side reference.

The LDO is now designed to be able to write a 4-bit word simultaneously, generating $10.8V - 10mA$ with a regulation rate $< 1\mu s$. In fact, when programming a cell, a current jump of $100ns - 2,5mA$ is obtained, the amplitude of which is limited by the saturation current of the $cell\#2$ selector transistor. The V_{prog} error must not exceed 5 % at 6-sigma. However, the presence of undershoot is not relevant as long as it is set within $1\mu s$. The design of a transient control circuit for overshoot / undershoot correction as in [3], [2] is not necessary.

The simplified LDO concept is shown in Fig. 4.2. V_{out} is connected to the drain of a PMOS pass transistor, M_8. Through a resistive feedback, made up of R_a and R_b, the output is fed back into the positive input of the Error Amplifier, EA, which compares it with the reference value, V_{ref} (bandgap

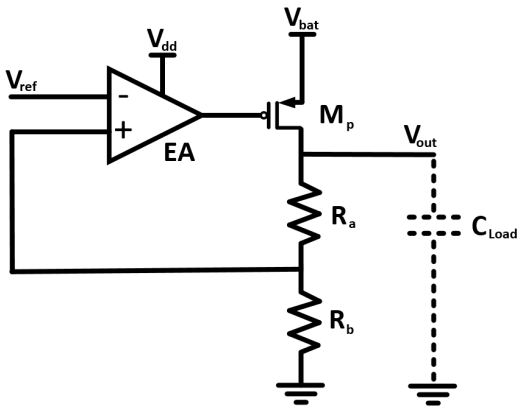

Figure 4.2: Typical LDO scheme

generated [27]). As the V_{out} decreases, the EA output will also decrease as it is in phase with the non-inverting input. The PMOS pass transistor then draws more current, causing V_{out} to rise.

The closed loop gain with respect to the battery voltage will be:

$$\frac{V_{out}}{V_{bat}} = \frac{A_2}{1 + A_1 \cdot A_2 \cdot \beta} \tag{4.1}$$

Where A_1 is the EA open loop gain, R_1 its output resistance, β is the feedback gain given by the resistor divider and A_2 is the common source open loop gain, the whose output resistance is R_2.

From Fig. 2.3, A_2 and β can be identified as:

$$A_2 = g_{m8} \cdot R_2 \tag{4.2}$$

$$\beta = \frac{R_b}{R_a + R_b} \tag{4.3}$$

where $R_2 = R_{ds8}//(R_a + R_b)//R_L$. While A_1 depends on the chosen EA. In any case, (4.1) will be less than 1. With a basic single-stage operational EA, the closed loop gain referred to the inverting input, V_{ref}, is:

$$\frac{V_{out}}{V_{ref}} = \frac{A_1 \cdot A_2}{1 + A_1 \cdot A_2 \cdot \beta} \tag{4.4}$$

$$A_1 = g_{mN} \cdot R_1 \tag{4.5}$$

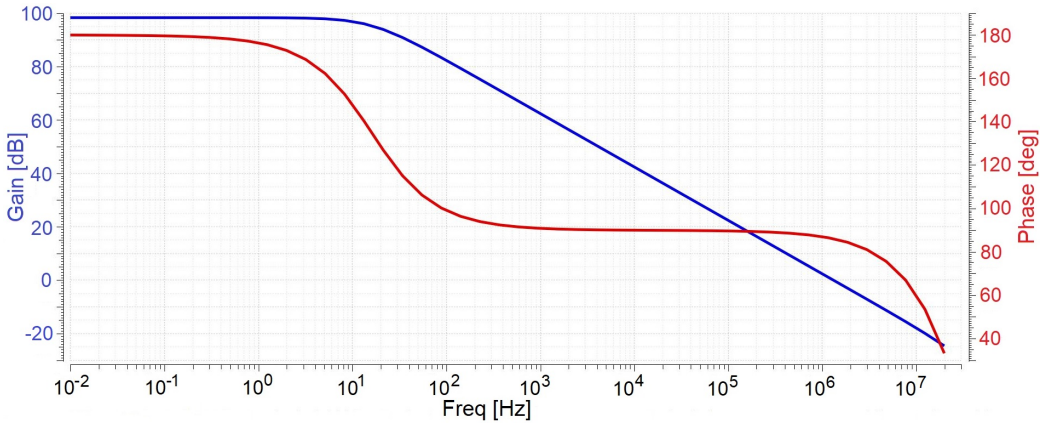

Figure 4.3: AC Gain and Phase

In a capacitor-less LDO, the pass transistor provides the dominant pole. The poles are located at:

$$P_1 \approx \frac{1}{R_1 \cdot C_1 + R_1 \cdot C_M \cdot g_{m8} \cdot R_2} \tag{4.6}$$

$$P_2 \approx \frac{g_{m8}}{C_1 + C_L} \tag{4.7}$$

The overall nominal frequency response of the loop gain, i.e. EA plus output stage with load of $R_L = (R_A + R_B)//R_L$, is shown in Fig. 4.3. The DC gain is about 98.4 dB and the unity gain bandwidth is about 1.35 MHz with a phase margin of 85 degrees.

Although no real Miller cap is adopted, the large M8 size introduces a large $C_{GD,8} = 4.2pF$, which acts as a Miller cap. $C_{GD,8}$ and $C_{GS,8} = 10.7pF$ identify the dominant pole in the EA output node: $1/(R_1 \cdot C_1)$, where:

$$R_1 = r_{DS,7} \parallel (r_{DS,2} \cdot gm_3 \cdot r_{DS,3}) \tag{4.8}$$

$$C_1 = C_{GS,8} + C_{GB,8} + A_2 \cdot C_{DG,8} \simeq A_2 \cdot C_{DG,8} \tag{4.9}$$

The non-dominant pole due to the output stage actually reduces to: $gm_8/(C_{DG,8} + C_{GS,8})$.

An additional Miller capacitor would result in a shift of the first pole to a lower frequency and, more importantly, a reduction in the slew rate,

negatively impacting the regulation speed during the transient. No output capacitors are added to improve stability.

Proper sizing of $M8$ is critical. A reduction in width of $M8$ would shift the dominant pole to the right as a result of a reduction of C_1. The smaller the device, the faster the adjustment will be. Meanwhile, gm_8 also shrinks, affecting DC gain and second pole position as well. The overall result is a deterioration of the PM, which ultimately leads to an undesirable overshoot. A higher V_{prog} would affect cell programming and should be avoided. Furthermore, the high frequency PSRR, a fairly important feature given the large line variations, would be negatively affected.

The output current is slightly increased by adding $R_L = 50k\Omega$ in parallel to the load ($R_a + R_b$ approximately 36k Ω) to increase gm_8. This improved stability and PSRR at the cost of negligible loss of gain.

The worst case for stability is when I_{load} is minimal, ie when the load is purely capacitive. Therefore, the OTP programming phase is not critical and the upper limit to the number of bit-cells programmed at a time is actually given by the maximum I_{load}, relative to the M_P dimensions.

4.2.1 Circuit schematic

The OTP circuit diagram is shown in Fig. 4.2, where the developed LDO block contains the diagram in Fig. 4.4. The LDO is composed of the cascade of an Error Amplifier and the pass-transistor output stage. The technology used allows to obtain 64dB of gain and 13.5MHz-UGB from a single EA stage. A more sophisticated EA would make the circuit bulkier with negligible benefits. To operate smoothly between two domains, LDO requires transistors with slightly different processes. The presence of a drift region between the drain terminal and the transistor channel transforms a medium voltage transistor into a high voltage one (represented with a thick gate in Fig. 4.4). M_1 and M_2 make up the EA input differential pair. However, M_3 and M_4 do not mirror the current in the two branches as usual. M_4 is in fact used to separate the high side from the low side, otherwise the M_2 transistor could be damaged by an exaggerated voltage drop on its terminals. In this way the drain of M_2 is limited above a quantity $V_{dd} - V_{th,M_4}$. Note that since the two domains are well separated, layout discrepancies are minimized. To make the differential input pair as accurate as possible under steady state conditions ($V_{ds\,M_1} \simeq V_{ds\,M_2}$), M_3 has been added. The current low side reference 5μ A, generated elsewhere, must be taken as a specification. Meanwhile,

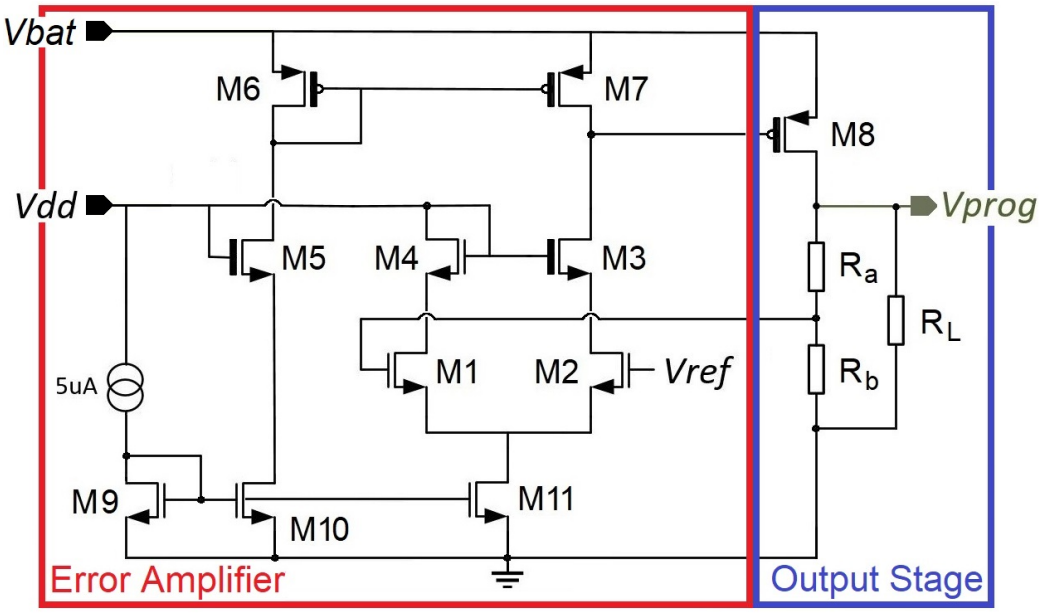

Figure 4.4: LDO schematic

the high side current mirror above forces a current of 30μA/2 via M_5 (the expected current in each branch). If the output voltage, V_{out}, drops, the feedback drops and the differential pair becomes unbalanced, with more current flowing on the M_2 branch. To restore the balance, $V_{ds\,M_2}$ decreases: this will open the M_P transistor pass more causing V_{out} to increase.

The OTP load is modeled as a capacitor, $C_L = 10pF$, 4 switched 5 kΩ resistors to simulate the programming of 4 cells and a piece-wise linear current source, I_{pwl}, to simulate the Fowler-Nordheim tunneling current. Indeed, as soon as the cells are connected to V_{prog}, a tunneling current begins to flow: at least $\sim$ 0.5mA per cell [1] (the minimum value of the load current is the worst condition for the stability).

All the device dimensions are resumed in Table 4.2.

Relevant LDO specifications are: dimming speed, idle output power and stability. Output line needs to adjust within 1μ s by minimizing overshoot and ringing. The output error at 6σ on the corners must be $\pm 5\%$.

In dual domain design, M3 and M5 are used to shield the low-side circuit from the high voltage section. M4 is added to make the V_{DS} of the input differential pair as similar as possible. In this way the drains of M1 and M2

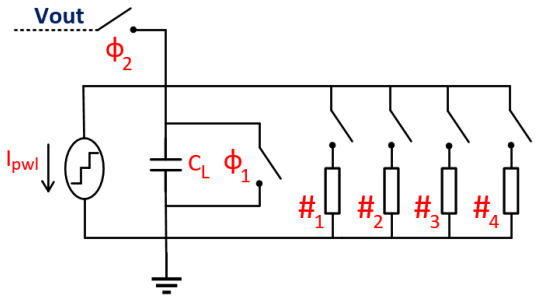

Figure 4.5: Model of the OTP load

Table 4.2: LDO transistors dimension

Name	**W/L [μm/μm]**	**Name**	**dim**
M1, M2	50/5	**Ra**	31.85 kΩ
M3, M4, M5	10/2	**Rb**	4.755 kΩ
M6	24/20	**RL**	50 kΩ
M7	48/20	I_{M9}	5 μA
M8	2300/2	I_{M11}	20 μA
M9	5/20	I_{M7}	10 μA
M10	5/20	$I_{M8,idle}$	500 μA
M11	20/20	V_{ref}	1.4 V

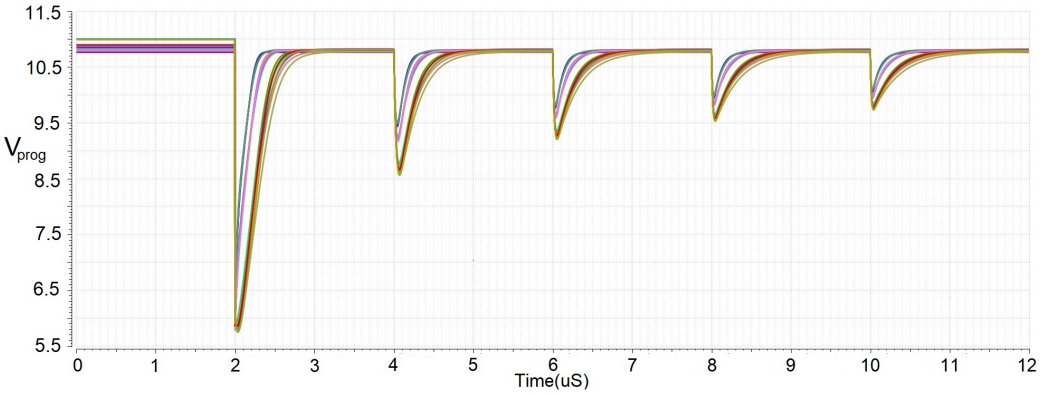

Figure 4.6: Transient simulation over Corners

have voltages $< 3V - V_{th} - V_{ov}$. Finally, the larger M1 and M2 sizes help reduce the offset of the input differential torque. M6 and M7 are designed to perform high output impedance, with a gate length of 20μ m. A longer channel would increase the C_{GS} capacity by reducing the PSRR. In fact, M6 and M7 impact in the PSRR is indeed relevant and must be minimized.

Each transistor that makes up the current mirrors has $L = 20\mu m$ to improve accuracy. The reference $5\mu A$, mirrored with a factor of 4, is pumped into the differential stage, meanwhile, through a high side mirror, $10\mu A$ is pumped from the above onto the non-inverting branch. With smaller mirror transistor lengths, the 5% to 6σ error is hardly realizable. The M8 pass transistor supplies $10mA$ to the OTP load plus $\simeq 500\mu A$ of idle current. The transistor always works in the saturation region as well as all other transistors in the scheme. M8 takes up $0.0046\mu m^2$, making it the most bulky device in the circuit.

Since the LDO V_{prog} must stabilize faster than $1\mu S$, to be as stable as possible when programming the cell, the SR^+ must be $\geq 10V/\mu s$. From this I_1 the current can be designed as:

$$SR^+ = \frac{I_1}{C_{GS,8}} \geq 10V/\mu s \rightarrow I_1 \geq 20\mu A \tag{4.10}$$

To ensure that the LDO adjusts within specification, worst case simulations were carried out. Overall 6-sigma on corners and MC simulations must be less than 5%. Fig. 4.6 shows the transient $Vprog$ line obtained by varying the processes together with temperatures from -40 to 175 degrees.

Simulations have shown that the system is robust and works well even in the worst conditions. The main source of error comes from the MC simulation. The overall positive and negative V_{prog} 6-sigma errors over corners and MC are $0.39V$ and $0.65V$, respectively, that is a positive error of 2.8% and a negative 3.7%. The same errors are obtained when $V_{bat} = 11\,V$. This ensures that under any conditions the circuit satisfies the $Vprog \pm 5\%$ target. The same errors are obtained in the worst case when $V_{bat} = 11\,V$, demonstrating that under any condition the circuit is able to adjust with an error of less than ±5%. Furthermore, to verify its reliability, a scan of up to 35V on the battery voltage was simulated. Due to the good line rejection, the positive and negative sigma are not really affected. In fact, the PSRR referred to V_{bat} (since $V_{prog}/V_{bat} = \frac{V_{prog}}{A_1 \cdot V_{ref}}$) should be [2]:

$$PSRR \approx \frac{R_1 + R_2}{A_1 \cdot R_2} \tag{4.11}$$

The V_{dd} PSRR is much better. Indeed the V_{dd} fluctuations don't really have an impact on the system (both input branches will move accordingly without appreciable changes). In DC, the $PSRR_{DC}$ would result $\approx -130dB$, at moderate-frequency [3] it would be $\approx -65dB$ and $\approx -45dB$ @100kHz.

As is clear, the LDO shows an undershoot when suddenly (ideal condition) the load is connected and a tunneling current begins to flow. Each time a cell is programmed, smaller undercuts appear. Since oxide breakdown is a cumulative process, degradation can even begin at lower voltages, such as those during undershoots, and reach its final breakdown at 10.8V.

Simulations are run to ensure proper operation, including mismatch between high side and low side devices. The proposed LDO performances are summarized in Table 4.3.

4.2.2 Experimental Results

The circuit prototype is made in 0.35μm CMOS technology. The top level schematic is shown in Fig. 4.7 while its layout is shown in Fig. 4.8. The total silicon area occupied by the chip is $0.06\,\text{mm}^2$. In the prototype, the LDO is realized together with two OTP loads in parallel (0.5 kB of memory, for a total equivalent capacitance of $10pF$) to validate its correct operations, collect some cell behaviors and verify their correspondence with the results of chapter 3. The resulting bonded silicon OTP reticle (2 x 0.7 mm^2) is

Table 4.3: Performance summary

Parameter	***Value***
Technology	350 nm
V_{bat}	13.5 V
I_{load}	10 mA
I_{load} rising time	100 nS
C_{load}	10 pF
Settling Time	1 μS
Loop Gain	83 dB
PSRR	-55 dB/1 kHz
Load Regulation	0.19 mV/mA
Line Regulation	0.64 mV/V
Area	0.06 mm^2

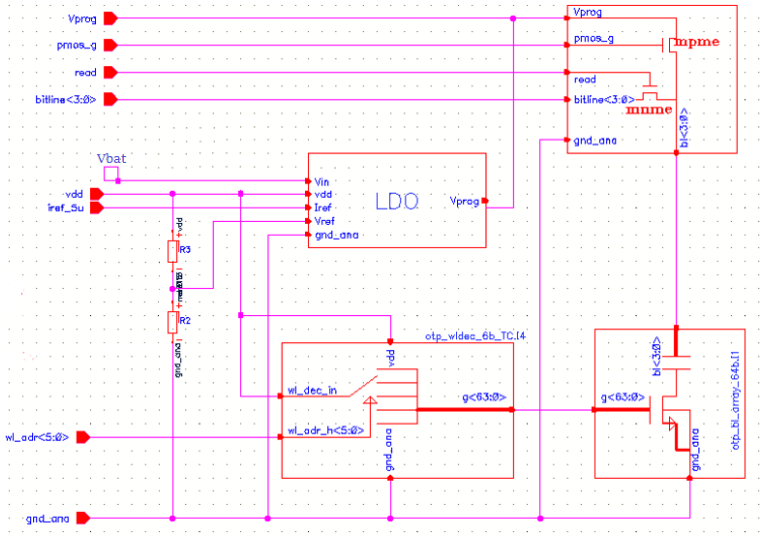

Figure 4.7: LDO circuit top level

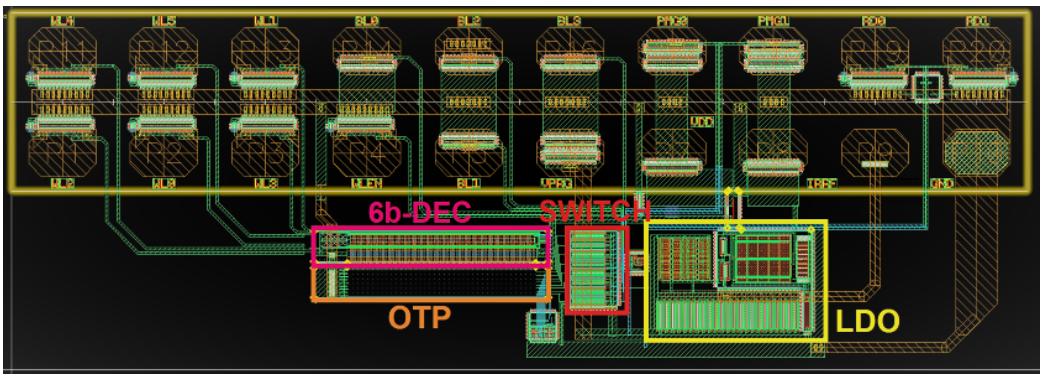

Figure 4.8: LDO circuit layout

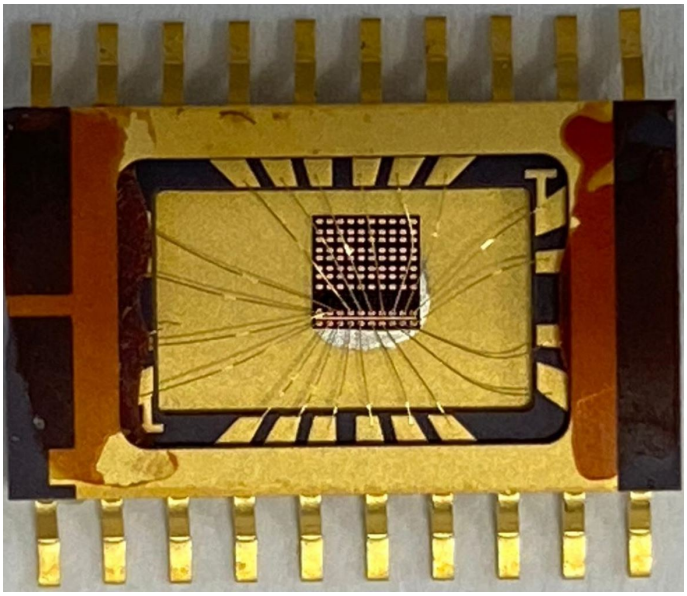

Figure 4.9: Ceramic package

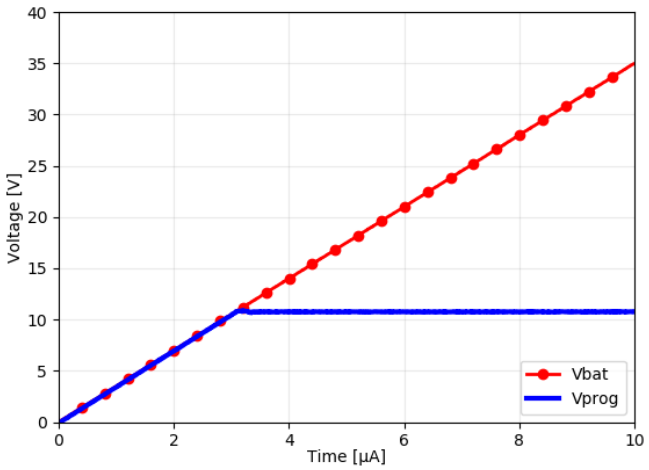

Figure 4.10: Vprog vs Vbat

shown in Fig. 4.9. Due to the reduced size, only the twenty 130 μm pads are distinguishable, as in Fig. 4.8.

The LDO is powered with a 3V power supply (which supplies $20\mu A$) and a nominal 13.5V power supply (which sinks the $500\mu A$ idle output current). In this development, for testing purposes, the references $5\mu A$ and V_{DD} are supplied externally, while the $Vref$ is produced internally with a resistive divider.

The waveform in Fig. 4.10, obtained with the oscilloscope, shows how the LDO maintains the programming voltage stably while ramping the battery voltage up to 35V, as for the cranking of the car.

As for how programmability works in on-chip OTP cells, several ceramic packaged chips are tested with the same procedure: (1) select a WL; (2) read the content to make sure the cells are virgin; (3) program them; (4) re-read to make sure the programming has taken place; (5) verify the goodness of the breakdown by listing the post-programming current. The procedure is repeated for different DUT temperatures: -40 to 150 degrees by means of a thermostream.

Fig. 4.11 plots the reading current flowing through the four OTP cells of the selected WL after programming. The nominal average is $\simeq 60\mu A$ for a suitably programmed cell (previously calculated by collecting thousands of silicon wafer data with the pico-probe station). Each point in Fig. 4.11 is the average of 4 reading currents. In total, the plot collects 100 data. The

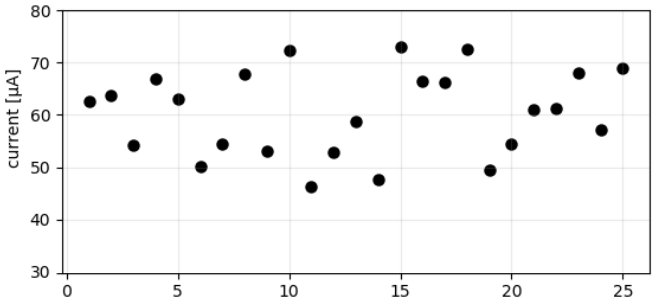

Figure 4.11: 100 post-programming currents

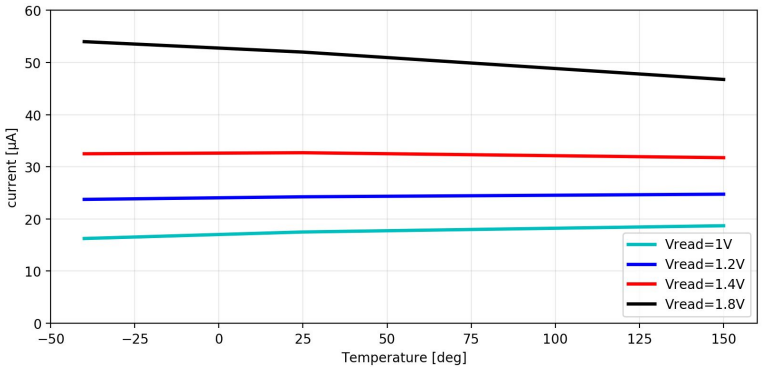

Figure 4.12: Read current changing temperature and Vprog

measured value $\sigma = 16\mu A$ is promising considering the statistical nature of the oxide breakdown event which returns a high diffuse current distribution. A complete OTP circuit possibly contains sense amplifiers for each BL to automatically read the written word, reading the current flowing in each BL and comparing it to a reference. In this application, the threshold is $30\mu A$.

The current read was also measured when changing the BL voltage. As expected, they are directly proportional, which means that the selector transistor works in its linear region and the post-break behavior of the oxide is similar to that of a resistor. The same procedure was repeated changing the DUT temperature with the thermostream. The trend of the current read with the temperature depends on the voltage V_{read}: for small voltages V_{read}, below $1.5V$, the behavior of the anti-fuse element dominates the selector transistor and the current grows with temperature. Above $1.5V$ the situation is reversed. Fig. 4.12 shows the results.

Eventually, a first glue of the failure rate derives from the oxide breakdown resistances. In particular, a cell in which at least $1\mu A$ of current flows through the cell when a $1.8VVread$ is supplied is considered well-programmed. However, a more meaningful statistic would be needed to ensure a given programming error rate before considering if the chip is suitable for a product. For this reason an error detection block will also be added on the final version of the circuit. Although the name One Time Programmable suggests that memory cannot be reprogrammed, using parallel OTP modules it is possible to build multi-OTP memory. There is no need to change the entire OTP chip in case of re-trimming a particular circuit: just select another module.

Chapter 5

Other Activities

While the first OTP memory test chip was being produced, other minor circuits were designed, which will be part of the next, more complete and product-like test chip. Exponential DAC and Line Observers are among them. Alongside this design activity, the author contributed to scientific dissemination within the European project iDEV40.

5.1 Exponential DAC

The need for an exponential DAC arises from the impossibility of actually verifying what the bit cell failure rate is. For example: a typical failure rate value is 0.01 ppm, but how can you prove it? Trying to program billions of cells isn't really an option. Furthermore, this would not give the customer the ability to actually control the failure rate on their own. Another, more efficient method is to include in the chip a circuit that allows to detect failed programmed cell by checking their R_{OTP}, or, equivalently, the oxide voltage drop when a certain current is forced though it. This way, if the voltage drop is smaller than a reference, surely the cell is hardly programmed, otherwise the programming has failed. Same procedure can be reiterated with different currents internally produced covering a very wide range to check the goodness of the breakdown: if even the highest current produces a smaller drop than the reference, the programming was very successful. Those currents cannot be produced linearly or the difference between two following values would be constant and it would take a huge circuit to cover the entire range $[1\mu A, 30\mu A]$. This is the reason behind the choice of an "exponential" DAC:

by only using 5 selection bits, 32 currents values are obtained. In fact, 16 values are already enough to accurately sample the critical interval and grow higher and higher as desired. The other values are simply not used. Furthermore, these current values are decided externally via digital bit. This explains the name "Digital to Analog Converter" (DAC).

A given reference current is forced through the bit-cell. The voltage drop across the anti-fuse element is then compared to a threshold by the sense amplifier which tells if the cell is programmed or not. The experiments in Chapter 3 allowed to grossly determine the order of magnitude of R_{OTP}, despite the large standard deviation of the distribution. The nominal current used to read the cell contents will be $1\mu A$. All other DAC values are for failure rate check only. $R_{OTP,avg} \simeq 3k\Omega$, but a cell can still be considered well programmed up to $\simeq 100k\Omega$. This means that the sense amplifier threshold must be: $100k\Omega \cdot 1\mu A = 100mV$. If the voltage across the dielectric is higher than this value the cell either is virgin or, in the worst case, not well programmed. When the whole OTP module has been programmed, through the exponential DAC the memory cells are checked. The first reading will be with the highest current value: if the voltage drop is already below the threshold, the dielectric is hardly broken down, with a very small resistance. Then the forced current is reduced and the check is repeated until eventually the cells are no longer read as programmed. Statistic is collected. Of course, for the lowest DAC current value, many cells are expected to fail, but yet the error rate is to be considered for the $1\mu A$ goal.

For this application, the current varies between $160nA$ and $30\mu A$. The variations in the output current of the DAC are obtained by switching the branches of the circuit on and off. Each branch, in fact, is designed to modify the current mirror ratio. First, to design the DAC, the current range and selection bits must be identified. The current 16 points are defined as in Table 5.1. Half of the 32 obtainable values is not foreseen in the drawing, as it would take points with little relevance. The points sought are only those in Fig. 5.1, where the abscissa represents the hexadecimal value of the 5-bit selection combination.

The conceptual circuit is shown in Fig. 5.2, subsequently implemented on Virtuoso (Cadence) to evaluate its functioning, robustness and precision. $Iref = 10\mu A$ is an internal current reference, while Out is the DAC current produced by the subsequent mirroring which will be used later to read the cell contents or check the failure rate. $Sel < 4 : 0 >$ are the selection bits. The most important bits, number 4 and 5, turn on / off the mirror

Exa-decimal value	Binary value	Current (μA)
0	00000	0.16
1	00001	0.26
3	00011	0.43
7	00111	0.64
8	01000	1.08
9	01001	1.82
11	01011	2.94
15	01111	4.41
24	11000	7.44
25	11001	12.48
27	11011	20.08
31	11111	29.83

Table 5.1: Exponential DAC specifications

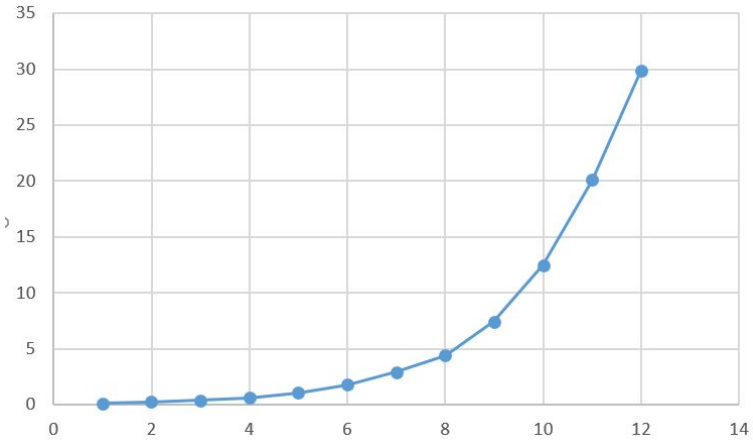

Figure 5.1: iDAC currents

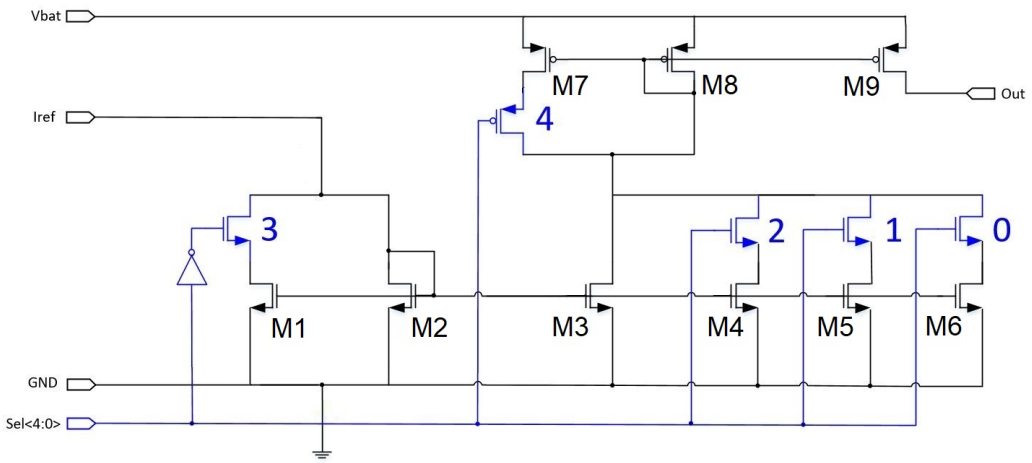

Figure 5.2: Exponential iDAC schematic

mothers below and above. A flip of those bits would result in huge current differences. The minimum bits, on the other hand, 0: 3, simply activate minor branches for small current steps. The behavior of the circuit is also understandable by looking at the dimensions of the device, listed in Table 5.2. When $Sel_<4:0>$ M1 is ON and most current flows think so. 1/6 of the remaining current flows through M2 and then mirrored by M3. Before arriving at the exit, $I_M3 = I_M7$ is further reduced by a factor of 6.

When one of the less important bits flips, $I_M3 \neq I_M7$, because $I_M7 = I_M3 + I_M4 + I_M5 + I_M6$, resulting in a smaller current increase. In fact, from the 4th and 5th bit higher steps derive: by turning off the mirror mothers, a factor of 6 is gained (as visible from Table 5.1 for 00000, 01000 and 11000) since both the input and output mirrors 1 to 1 mirror.

Corners and Montecarlo simulations are carried out to verify the robustness and accuracy of the circuit. In this way it is possible to estimate the 6-sigma deviation of the currents from their nominal values of the Table 5.1. However, the most important error to check is the one corresponding to $1\mu A$: 8% a 6-sigma on angles and MC. Higher 6-sigma errors are obtained for smaller currents, an vice versa for for higher currents.

Device	W (μm)	L (μm)
M1	12	10
M2	2	10
M3	1.5	10
M4	1	10
M5	1.5	10
M6	2	10
M7	72	10
M8	12	10
M9	12	10

Table 5.2: iDAC Transistor sizes

5.2 Line Observers

Line observers are used to check BL and WL connections. Detecting possible shorts or open circuits between adjacent BL or WL is a wise technique to prevent unexpected multiple failures during the reading phase. The concept of source observer is shown in Fig. 5.3: if the ground connection is ensured, it is not possible that the line will be pulled to VDD with only 500 nA. So $WL\_Source\_OK = 1$. It is sufficient that a connection is interrupted, that the line goes up to VDD and the corresponding output transistor turns on, returning an error flag $WL\_Source\_OK = 0$.

The situation changes slightly for the WL observer. Contrary to Source connections, shorting with adjacent lines is problematic (as they are no longer connected to ground). In fact, a circuit like the one in Fig. 5.3 is able to detect only line interruptions, but not a short circuit. Since the WL controls the gate of the selector transistor, it is important to have a short circuit detection or there may be one / more incorrect programming. The correct concept would be that of Fig. 5.4. First the even WLs are brought to VDD and the odd ones connected to GND. Similarly to the source observer, a 500nA tries to break down the even WL, but this does not happen until there are no interruptions, so the output PMOS are open and the node "$out1$" is 0. The odd WLs are instead set to ground, which means that even if there is indeed an open circuit, it will not be detected due to the pull-down current. So, "$out2$ is 1. The XOR gate of the two signals would give a 1: everything is fine. As soon as a short occurs, the even WL would be brought to zero, producing $out1 = 1$. In that case , the XOR gives 0: something is

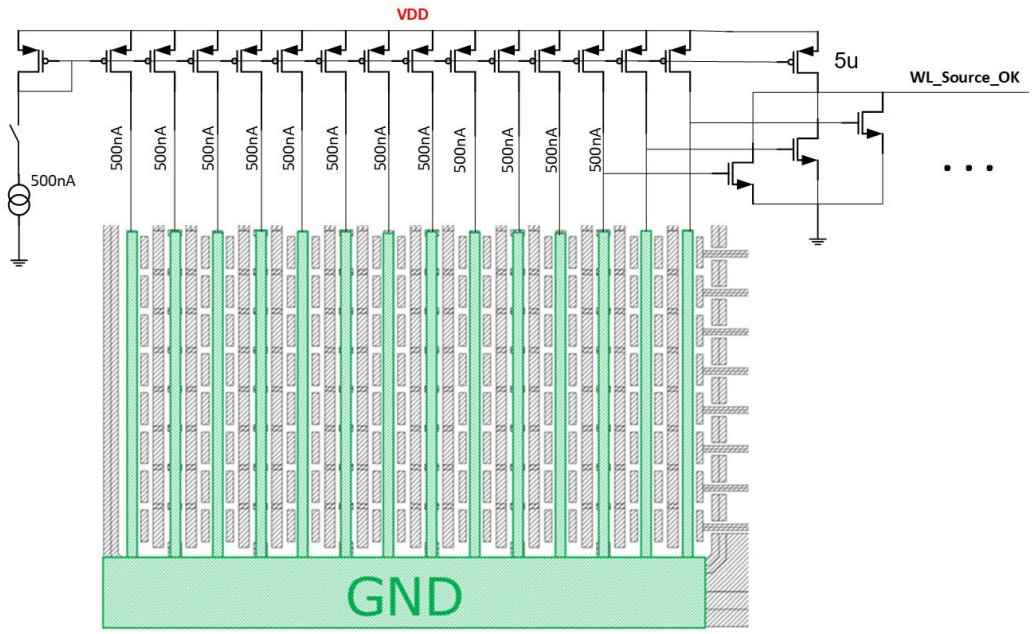

Figure 5.3: Source Observer concept Schematic

wrong. Also, if an even WL is broken, the line would be grounded again by the current of 500 nA, producing an error flag. However, a break in the odd WL is not yet detectable: the odd WL would still be zeroed, giving $out2 = 1$. One solution is to repeat the procedure one more time reversing the bias of even and odd WLs. Now, if everything is correct, $out1 = 1$ and $out2 = 0$ give an ending 1. During this phase an odd WL interrupted would be detected as it is expected to be high, while, in the open circuit, it is actually grounded .

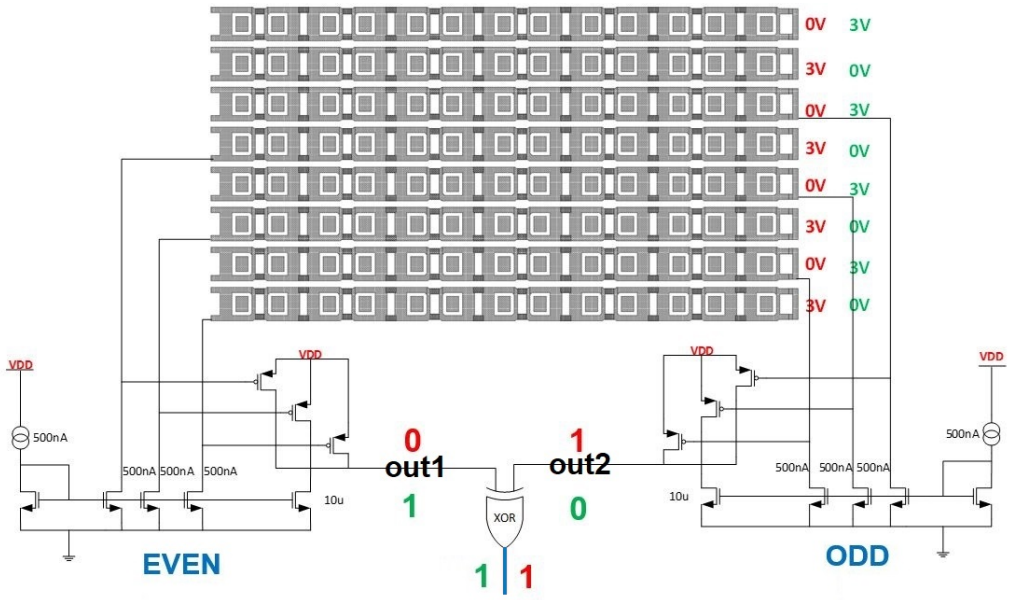

Figure 5.4: WL Observer concept schematic

5.3 iDev40 European Project

UNIMIB and Infineon Technologies are both partners of the European Integrated Development 4.0 project (iDEV40) committed to supporting and implementing digitization strategies for the 4.0 semiconductor industries. The author was part of the iDEV40 community, working on "Work Package 3 - Use Case 15", focused on digital transformation in semiconductor manufacturing using as an application case the development methodology of a monolithic DC / DC converter for automotive applications, object of parallel studies to the main OTP topic. All information and details relating to the topics and works of iDEV40 are available in the literature [17], [18], [19], [20], [21], [22], [23], [24], [25], [26].

A part of the work has been performed in the project iDev40. The iDev40 project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 783163. The JU receives support from the European Union's Horizon 2020 research and innovation programme. It is co-funded by the consortium members, grants from Austria, Germany, Belgium, Italy, Spain and Romania. The information and results set out in this publication are those of the authors and do

not necessarily reflect the opinion of the ECSEL Joint Undertaking.

Chapter 6

Conclusions

The work retraces all the phases faced in the realization of the OTP memory using a high-power technology. After the general introduction of OTP memory, the theory and physical models underlying the oxide breakdown are presented. Indeed, to obtain a product-like version of an OTP memory, it is of primary importance to understand the behavior of the oxide and define the circuit design specifications. The theory is applied to Infineon 350nm CMOS power technology, to study the dielectric before its implementation on the OTP memory cell. Experiments are also carried out on the selector transistor: characterization aimed at finding the best efficiency-price ratio, where the efficiency derives from the programming of the oxide, while the price is correlated to the size of the bit cell. The physical models that describe the dielectric breakdown event can be different, depending on the characteristics of the oxide, the mechanisms of generation of the defect, etc. Each physical model inevitably leads to a different trend of the breaking time. A time-dependent dielectric breakdown (TDDB) technique is proposed for an OTP bit cell composed of a transistor and a capacitor, where the capacitor represents the anti-fuse element. The aim is to find the correlation between the T_{BD}, the capacitor area and the applied voltage in order to optimize the OTP cell in terms of programming power, memory density and speed or, vice versa, for the device failure analysis to understand how to maximize T_{BD}. In fact, the same procedure could be used elsewhere for predicting device life, rather than being used to determine the best anti-fuse programming conditions.

The characterization is conducted for a 7.7 nm and a 2.2 nm oxide layer. In the latter case, the experimental results show how the modeling of the break becomes really challenging. However, the relevance of the wear and time to failure pattern has been demonstrated. In addition, a dielectric breakdown strength was estimated. All the extrapolated parameters serve as specifications for the actual

OTP memory design. A memory chip test was designed, simulated and subsequently tested in a ceramic package. The core of the chip is the low dropout regulator which consistently and accurately supplies the programming voltage to the OTP module. The 10.8 V dual domain LDO project was described. The simple yet thoughtful design avoided using bulky transient control circuits to minimize undershoot, resulting in a reliable and competitive circuit.

The lab experiments eventually proved the test chip's functionality even with drastic battery sweeping and temperature changes. In this way, the correct choice of the oxide was also demonstrated.

Peripheral circuits were also designed to build the second, product-like test chip, which is expected to be delivered in March 2022. The final operation will then be verified, possibly bringing to the market an OTP anti-fuse memory for automotive applications.

Bibliography

[1] S.A. Lombardo, K.L.Pey, J. Stathis, F. Palumbo, "Dielectric Breakdown mechanisms in gate oxides", Journal of Applied Physics, Dec. 2005

[2] D. J. DiMaria, E. Cartier "Mechanism for stress-induced leakage currents in thin silicon dioxide films", J. Appl. Phys., Vol. 78, No. 6, 15 Sept. 1995

[3] Kalus F. Schuegraf, Chenming Hu, "Hole Injection SiO2 Breadkown Model for Very Low Voltage Lifetime Extrapolation", IEEE transaction on electronic devices, Vol 41, No.5, May 1994

[4] Muhammed A. Alam, Jeff Bude, Andrea Ghetti, "Field Acceleration For Oxide Breakdown – Can An Accurate Anode Hole Injection Model Resolve the E vs. 1/E Controversy?", IEEE 38th Annual International Reliability Physics Symposium, San Jose, California, 2000

[5] A. Shluger, "Defects in Oxides in Electronic Devices", Handbook of Materials Modeling, Springer, 2019

[6] C. H. Yang, S. C. Chen, Y. S. Tsai, R. Lu, and Y.-H. Lee, "The Physical Explanation of TDDB Power Law Lifetime Model Through Oxygen Vacancy Trap Investigations in HKMG NMOS FinFET Devices", IEEE, 2017

[7] L. Fengming, S. Jiang, L. Xiaoyu, W. Xianghao, "Validation Test Method of TDDB Physics-of-Failure Models", Prognostics & System Health Management Conference, 2012

[8] K.F. Schuegraf, C. Hu, "Hole injection oxide breakdown model for very low voltage lifetime extrapolation", 31st Annual Proceedings Reliability Physics, Atlanta, USA, 1993

[9] Muhammad A. Alam, B. E. Weir, P. J. Silverman, Y. Ma, and D. Hwang. "The Statistical Distribution of Percolation Resistance as a Probe into the Mechanics of Ultra-thin Oxide Breakdown"

[10] O.Gasparri, M.Bernardoni, P. Del Croce et al. "Ultra-thin oxide breakdown for OTP development in power technologies". Elektrotech. Inftech. 138, 44–47 (2021). https://doi.org/10.1007/s00502-020-00838-1

[11] J. Pèrez-Bailòon, A.Màrquez, B. Calvo, "Transient-enhanced Output-Capacitorless CMOS LDO Regulator for Battery-operated Systems" 2017 IEEE ISCAS, 2017, PP 1-4

[12] S. Heng, C.K. Pham "A low-Power High-PSRR Low-Dropout regulator With Bulk-gate Controlled Circuit", IEEE Transactions on Circuits and Systems -II, Vol.57, No.4, April 2020

[13] M. Khan, M. H. Chowdhury, "Capacitor-less Low-Dropout Regulator (LDO) with Improved PSRR and Enhanced Slew-Rate", 2018 IEEE ISCAS, Florence, Italy, 2018, pp. 1-5, doi: 10.1109/ISCAS.2018.8351039.

[14] E. Miranda , "Method for extracting series resistance in MOS devices using Fowler-Nordheim plot", Electronic Letters, vol.40, pp 1153-1154, Sept. 2004

[15] J. Stathis, "Physical and predictive models of ultrathin oxide reliability in CMOS devices and circuits", Device and Materials Reliability, IEEE, vol.1, no.1, pp. 43-59, Mar 2001

[16] M. Deloge, "Analysis of ultrathin gate-oxide breakdown mechanisms and applications to antifuse memories fabricated in advanced CMOS processes", INSA de Lyon, chap. 5, 2011

[17] O. Gasparri, R. Di Lorenzo, P. Del Croce, A. Baschirotto, "PCMC DC-DC Converter Development Methodology by Means of dSPACE", Digital Transformation in Semiconductor Manufacturing, 136-144

[18] O. Gasparri, A. Baschirotto, P. Del Croce, A. Pidutti, "Power converter control using calculated average current", US Patent 10,433,378

[19] R Di Lorenzo, O. Gasparri, A. Pidutti, P. Del Croce, A. Baschirotto, "On-Chip Power Stage and Gate Driver for Fast Switching Applications", 2019 26th IEEE International Conference on Electronics, Circuits and Systems 2019

[20] O. Gasparri, R. Di Lorenzo, P. Del Croce, A. Pidutti, A. Baschirotto, "Variable off-Time Peak Current Mode Control (VoT-PCMC) as method for average current regulation in Buck Converter Drivers", 2019 26th IEEE International Conference on Electronics, Circuits and Systems 2019

[21] O. Gasparri, P. Del Croce, A. Baschirotto, "DC-DC Buck Converter Driver with Variable Off-Time Peak Current Mode Control" https://astesj.com/v05/i06/p42/ 5 (6), 347-352, 2020

[22] O. Gasparri, R. Di Lorenzo, A. Pidutti, P. Del Croce, A. Baschirotto, "DC-DC Buck Converter with Constant Off-Time Peak Current Mode Control" 2020 27th IEEE International Conference on Electronics, Circuits and Systems, 2020

[23] O. Gasparri, M. Bernardoni, P. Del Croce, A. Baschirotto, "Ultra-thin oxide breakdown for OTP development in power technologies", e & i Elektrotechnik und Informationstechnik 138 (1), 44-47, 2021

[24] O. Gasparri, A. Pidutti, P. Del Croce, A. Baschirotto, "A Fast Switching Current Controlled DC/DC Converter for Automotive Applications", Journal of Electrical and Electronic Engineering 9 (4), 123-128, 2021

[25] O. Gasparri, A Bozic, P. Del Croce, A. Baschirotto, "High-Voltage Double-Domain Low-Dropout regulator for fast varying output loads", IEEE ICECS 2021

[26] O. Gasparri, A Bozic, P. Del Croce, A. Baschirotto, "A Low-Dropout Regulator for One Time Programmable(OTP) Memories in Automotive Applications", IEEE ICECS 2021

[27] E. Barteselli; L. Sant; R. Gaggl; A. Baschirotto, "Design Techniques for Low-Power and Low-Voltage Bandgaps", Electricity 2021, 2, 271-284. https://doi.org/10.3390/electricity2030016

[28] E. Barteselli, L. Sant, R. Gaggl and A. Baschirotto, "A First Order-Curvature Compensation 5ppm/degC Low-Voltage & High PSR 65nm-CMOS Bandgap Reference with one-point 4-bits Trimming Resistor", SMACD / PRIME 2021; International Conference on SMACD and 16th Conference on PRIME, 2021, pp. 1-4