energy-driven devices and circuit designs for resistive random access memories

Paola Trotti

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : PHYS - Physique
Spécialité : Physique appliquée
Unité de recherche : Laboratoire d'Electronique et de Technologie de l'Information (LETI - CEA)

Dispositifs et circuits à convenance énergétique pour les mémoires résistives

energy-driven devices and circuit designs for resistive random access memories

Présentée par :

Paola TROTTI

Direction de thèse :

Gaël PILLONNET Ingénieur HDR, CEA Centre de Grenoble	Directeur de thèse
Gabriel MOLAS CEA	Co-encadrant de thèse
SAMI OUKASSI Ingénieur Chercheur, CEA-leti Grenoble	Co-encadrant de thèse

Rapporteurs :

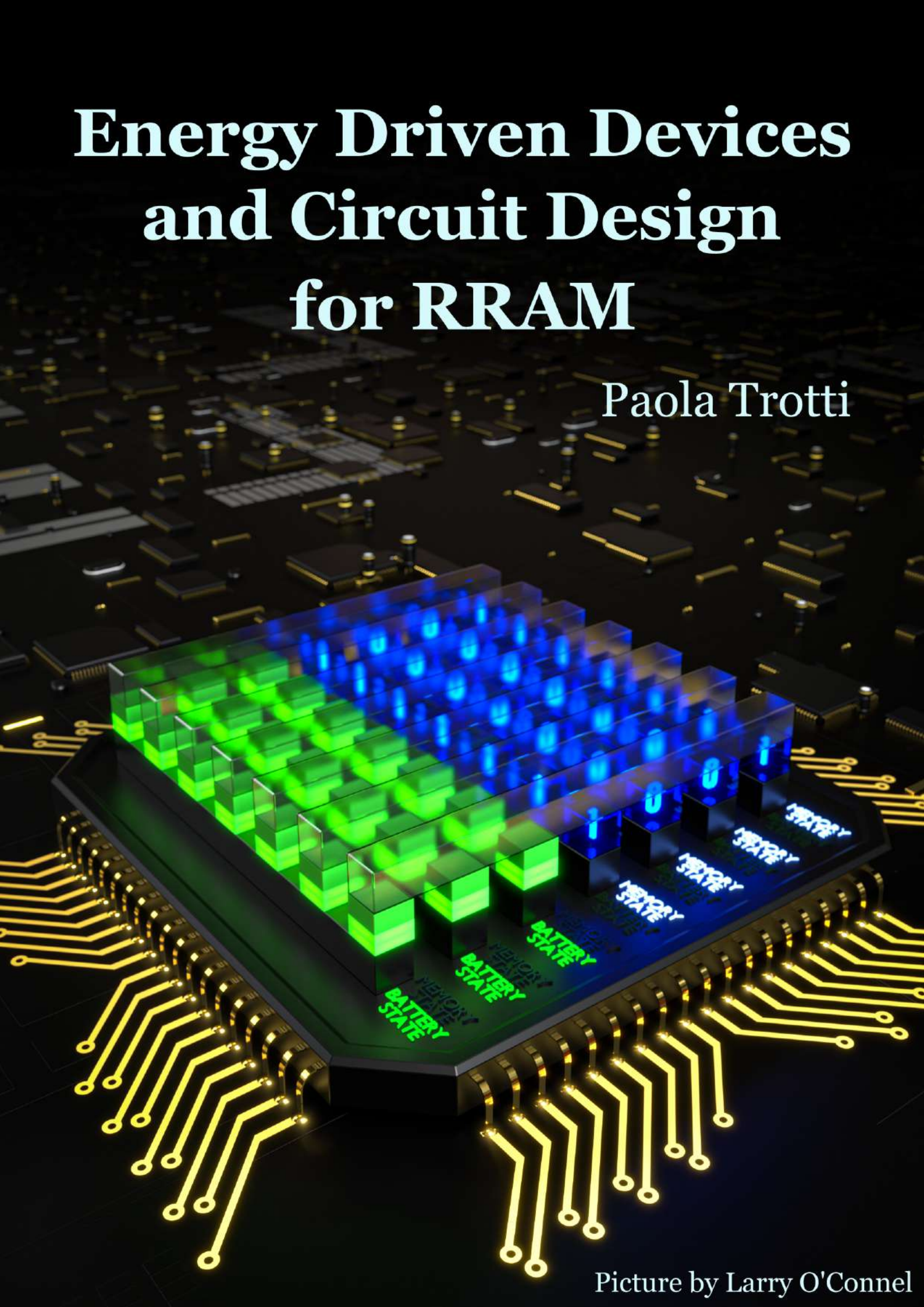
MARC BOCQUET
Professeur des Universités, AIX-MARSEILLE UNIVERSITE
ABDELKADER SOUIFI
Professeur des Universités, INSA LYON

Thèse soutenue publiquement le **8 décembre 2022**, devant le jury composé de :

MARC BOCQUET Professeur des Universités, AIX-MARSEILLE UNIVERSITE	Rapporteur
ABDELKADER SOUIFI Professeur des Universités, INSA LYON	Président
QUENTIN RAFHAY Maître de conférences HDR, GRENOBLE INP	Examinateur
LIONEL TORRES Professeur des Universités, UNIVERSITE DE MONTPELLIER	Examinateur
CLAIRE FENOUILLET-BERANGER Ingénieur HDR, CEA CENTRE DE GRENOBLE	Examinatrice
ILIA VALOV Docteur en sciences, Forschungszentrum Jülich	Examinateur

Energy Driven Devices and Circuit Design for RRAM

Paola Trotti



Picture by Larry O'Connel

ABSTRACT

ENERGY-DRIVEN DEVICES AND CIRCUIT DESIGNS FOR RESISTIVE RANDOM ACCESS MEMORIES

Nowadays, electronic devices are implemented to carry out a wide set of tasks, ranging from high performance processing to low power sensing. Aggressive technological downscaling has allowed to obtain smarter devices, where increased design complexity brings a higher number of functions per chip area, along with faster operational speed and lower power consumption. However, advanced technological nodes have been suffering from architectural limitations, such as the Von Neumann's bottleneck, as well as delay and power loss over interconnections lines. As a results, new solutions are demanded in order to achieve ever increasing performances. To tackle the challenge, the so-called *More-than-Moore* devices have been emerging, where the hardware architecture as a whole is revisited, for example as in In Memory Computing (IMC) or Near Memory Computing (NMC), where the computation is brought onto, or in close proximity to the memory.
This thesis presents alternative solutions to the need of both process miniaturization and better energy efficiency in the field of semiconductor memories. It focuses on emerging Resistive Random Access Memory (RRAM) technology, and consists of two main conceptual parts:

- △ **The experimental study on RRAM as energy source**, where an architecture able to retain both information and energy is envisaged. Cyclic Voltammetry tests are performed on a wide range of State-Of-the-Art (SOA) devices, evaluating their electrochemical properties. Major traits of RRAM as energy source are outlines and compared to SOA alternatives. Promising energy and power densities are derived, and prospective implementation fields are discussed.
- △ **Novel, energy-efficient circuit designs for SOA RRAM write/erase operations.** Two different approaches are presented: a programming scheme where the set energy is stored on a charged capacitor, and a Current Digital to Analog Converter (IDAC) based architecture. Design details in 130nm CMOS technology are presented, where the memory array is integrated as BEOF process. We deliver proof of concept, and demonstrate that in both cases a clear advantage in terms of energy cost can be achieved over the standard pulsed-voltage method.

KEYWORDS: RRAM, Dual-behaviour memory, Energy-efficient memory design, Capacitive Resistive Switching, CQS, DAC based RRAM programming.

RÉSUMÉ EN FRANÇAIS

DISPOSITIFS ET CIRCUITS À CONVENANCE ÉNERGÉTIQUE POUR LES MÉMOIRES RÉSISTIVES

De nos jours, les appareils électroniques sont mis en œuvre pour effectuer un large éventail de tâches, allant du traitement haute performance à la détection de faible puissance. Une réduction d'échelle technologique agressive a permis d'obtenir des appareils plus intelligents, où la complexité accrue de la conception apporte un plus grand nombre de fonctions par zone de puce, ainsi qu'une vitesse opérationnelle plus rapide et une consommation d'énergie plus faible. Cependant, les nœuds technologiques avancés ont souffert de limitations architecturales, telles que le *Von Neumann Bottleneck*, ainsi que de retards et de pertes de puissance sur les lignes d'interconnexion. En conséquence, de nouvelles solutions sont demandées afin d'atteindre des performances toujours plus élevées. Pour relever le défi, les appareils dits *More-than-Moore* ont émergé, et l'architecture matérielle dans son ensemble a été revisitée, comme avec l'*In Memory Computing (IMC)* ou *Near Memory Computing (NMC).*
Cette thèse présente des solutions alternatives au besoin de miniaturisation des processus et d'une meilleure efficacité énergétique dans le domaine des mémoires semi-conductrices. Elle se concentre sur la technologie émergente de mémoire vive résistive (RRAM) et se compose de deux parties conceptuelles principales :

- △ **L'étude expérimentale sur les RRAM comme source d'énergie**, où une architecture capable de retenir à la fois l'information et l'énergie est envisagée. Les tests de Voltammetrie Cyclique sont effectués sur une large gamme de memoires, évaluant leurs propriétés électrochimiques. Les principales caractéristiques des RRAM en tant que source d'énergie sont présentées et comparées aux alternatives. Des densités d'énergie et de puissance prometteuses sont extrapolées et les domaines de mise en œuvre potentiels sont discutés.
- △ **Conceptions de circuits économes en énergie pour les opérations d'écriture des memoires RRAM.** Deux approches différentes sont présentées : un schéma de programmation où l'énergie est stockée dans un condensateur chargé, et une architecture basée sur un convertisseur digital-analogique en courant (IDAC). Les détails de conception en technologie CMOS 130 nm sont présentés. Dans les deux cas, un avantage clair en termes de coût énergétique peut être obtenu par rapport à la méthode standard de programmation en tension pulsée.

MOTS-CLÉS: RRAM, Mémoire à double comportement, conception de mémoire économe en énergie, Memoires Resistives, commutation résistive capacitive, CQS, programmation des RRAM basée sur DAC.

An expert is someone who has made all of the possible mistakes in a very narrow field of study. — Niels Bohr

ACKNOWLEDGEMENTS

Many people deserve to be thanked for the completion of this thesis work.
First of all, my gratitude goes to my thesis director, Gaël Pillonnet, and my supervisors, Gabriel Molas and Sami Oukassi, for the opportunity of conducting my research, and for being a persistent source of support, both technical and moral. I extend my thanks the people in CEA-Leti who actively helped in my work, in particular: Yasser Moursy, Niccolo Castellani, Olivier Billoint, Carlo Cagli, Mathieu Aubras, Larry O'Connell, Gabriele Navarro, Philippe Blaise, Etienne Nowak and Stéphanie Robinet.
I thank my friends, with whom I shared precious moments during this jurney; a special mention goes to Sota Sawaguchi, Leo Laborie, Benjamin Bonnard, Luis Cubero Montealegre, Giusy Lama, Sergio Dominguez and Alessandro Bricalli.
Finally, to my family, whose unwavering support has kept me motivated during the most challenging times.

Paola Trotti
October 2022

CONTENTS

ACRONYMS

SOA State Of the Art

RAM Random Access Memory

SCM Storage Class Memory

NVM Non-Volatile Memory

RRAM Resistive Random Access Memory

CBRAM Conductive Bridge Random Access Memory

LRS Low Resistive State

HRS High Resistive State

DUT Device Under Test

TEM Transmission Electron Microscopy

CV Cyclic Voltammetry

SMU Source Monitor Unit

CQS Constant Charge Source

CVS Constant Voltage Source

DAC Digital to Analog Converter

WT Write Termination

WV Write Verify

Part I

INTRODUCTION

1 INTRODUCTION

1.1 CONTEXT

In the past, technological development has been majorly pushed by transistor downscaling [1, 2]; however, as CMOS nodes have advanced, the performance bottleneck has shifted towards interconnects [3–5] and architectures. *The Memory Wall* [6–8] is a well-known problem, where the widening throughput disparity between processor and memory demands a re-visitation of the standard computer architecture. As a result, the so-called *More-than-Moore* devices have been emerging, such as In Memory Computing (IMC) [9–13] or Near Memory Computing (NMC) [8, 14–17], where the computation takes place either onto the memory itself or in its close vicinity.
This chapter gives context to our research, aimed to provide an alternative solution to the need of process miniaturization, reduced latency and better energy efficiency, in the field of semiconductor memories. Various emerging memories are presented, with major focus on Resistive Random Access Memory (RRAM), the technology adopted in our study. Its working mechanism is explained and basic design concepts are illustrated.

1.2 SEMICONDUCTOR MEMORIES

A computer memory can be defined as a piece of hardware whose task is to retain information, either temporarily of permanently, encoded through binary digits (bits) [18, 19]. It can be further categorized into *main* or *primary*, and *auxiliary* or *secondary* memory; the former generally being responsible of holding data and instructions during a program's execution, while the latter delivers long-term storage [19, 20].
Figure 1 shows a chart reporting the main memory types implemented in today's computer architectures. A further distinction can be made depending on whether the stored data is lost after the system's power off: is this case, the memory is called *volatile*, and *non-volatile* otherwise [18, 19, 21]. Examples of volatile technologies are SRAM (Static Random Access Memory) [22–25] and DRAM (Dynamic Random Access Memory) [22, 26, 27], while Read Only Memories (ROM) [28], Flash [29–31] and Hard Disks [32] are non-volatile.
In Figure 2a is reported a schematic view of Von Neumann's architecture [12, 33, 34]: the most largely adapted structure of today's digital computers [9, 34, 35]. At its core there is a Central Processing Unit (CPU), or processor, which executes the instructions that constitute a computer program; it comprises of a Control Unit (CU), which directs the processor operations, as well as an Arithmetic-Logic Unit (ALU), which carries out the actual calculation. The primary memory is highlighted: its role is to supply a program's instructions and data to the CPU, as well as to store the processed data. Hence, the communication between the

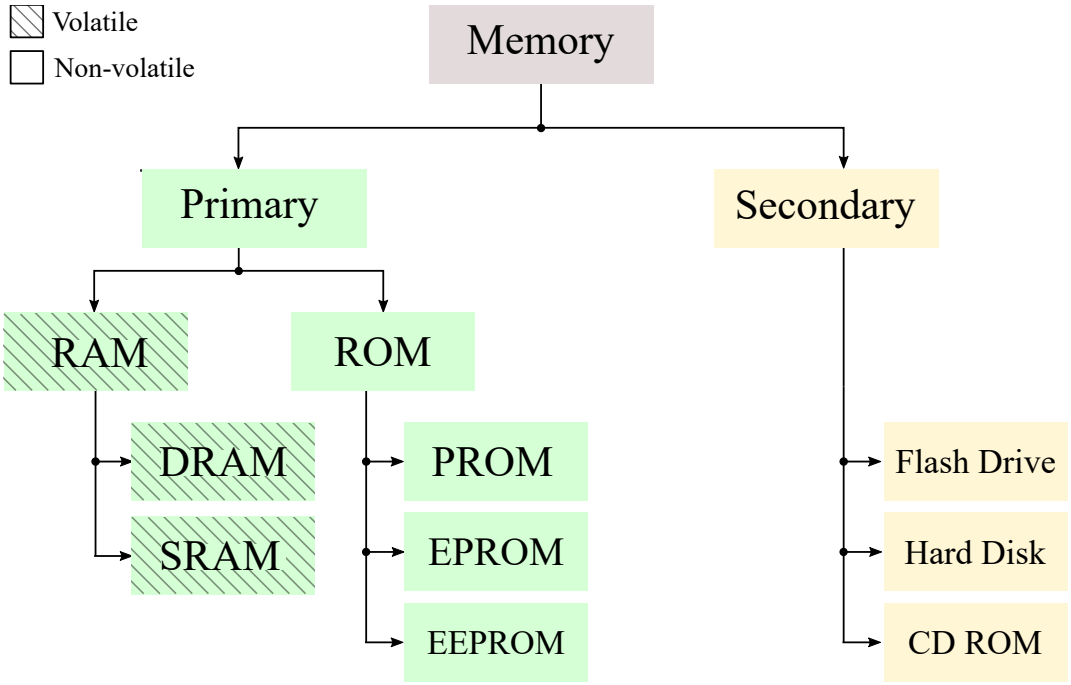


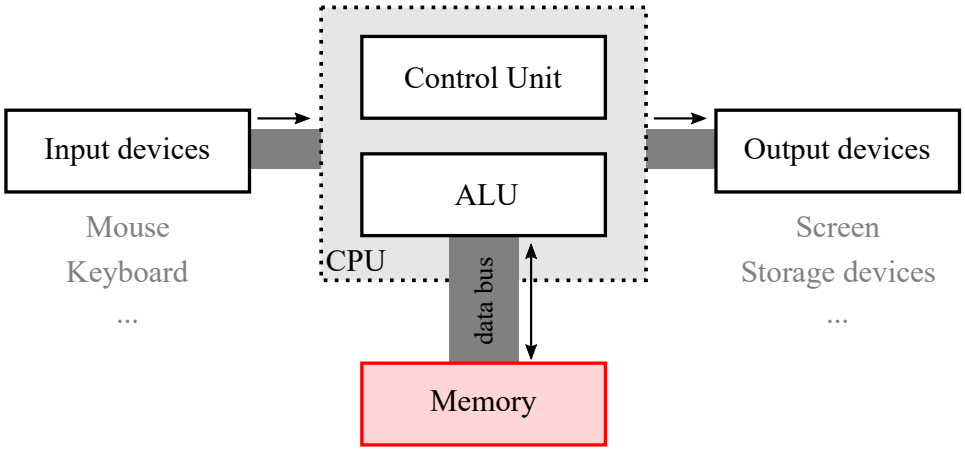
Figure 1 – Memory classification: primary and secondary storage, with main SOA technologies. Adapted from [20].

primary memory and the processor is bidirectional, and takes place over the data bus.

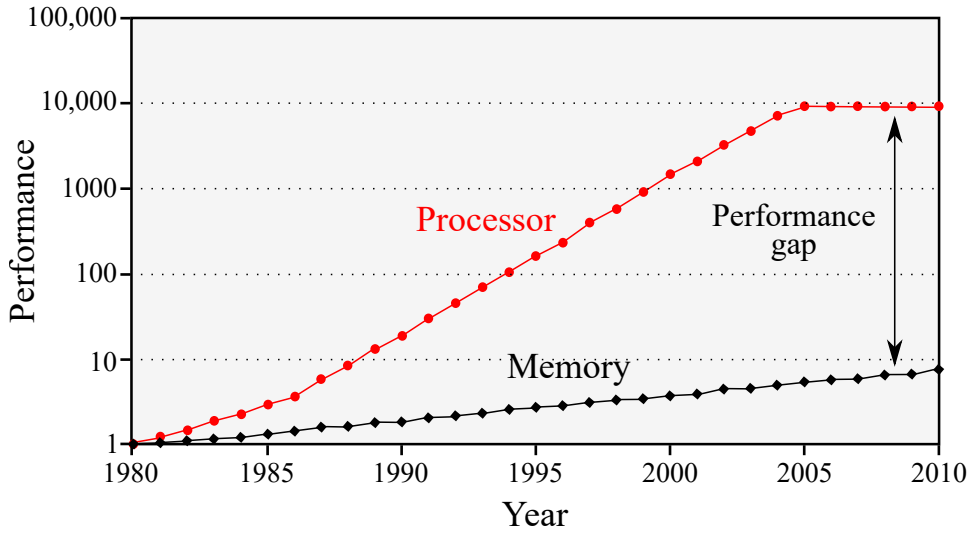
1.2.1 *The Memory Hierarchy*

Ideally, the speed at which the processor elaborates the incoming data matches the pace at which the the memory delivers it, so that no time is lost in waiting. However, a widening performance gap has been observed over the years, where the processor throughput [1] has been increasing at a faster pace compared to the memory access time [36]. As a result, the memory latency, i.e. the amount of time that passes between a request for data and the moment the data is available for external use, has become a performance bottleneck [19, 35, 37]. This problem has been known as the *Memory Wall* [6, 7, 38] or the *Von Neumann's bottleneck* [8, 39], and it is illustrated in Figure 2b, which reports the evolution of the processor and primary memory speed over the year.
Recent memory designs have been focusing on reducing power consumption as well as access times; this is due to the fact that enlarged, fast memories will dissipate considerable amount of energy either through dynamic [2] and static losses [3] [7].
The ultimate memory would be fast, cheap, highly scalable, low energy-consuming, and non-volatile. However, in reality these benefits trade with each other; conse-

1. In the CPU context, the throughput expresses the number of instructions that can be executed in a time unit.
2. Dynamic power is consumed when the memory is written/read.
3. Static consumption derives from leakage when the memory is not operated.



(a) Von Neumann's architecture, where the (primary) memory is highlighted. The data bus connects the CPU to the memory, which stores program instructions and data.



(b) The growing gap between the processor and the memory performances. In red, the rate of CPU requests to the memory, while in black, the memory access time. In more recent years, the processor growth tends to flatten, as the path to higher performance started relying on multi-processing, rather than a single processor speed [7]. Redrawn and adapted from [7].

Figure 2

quently, modern computers employ multiple memory units, which differ by technology and architecture, in an effort to shrink the performance disparity between the processor and the memory [18–20]. Figure 3a reports the so-called *Memory Hierarchy* [19, 40], where each layer stands for a different technology: speed and cost decrease towards the bottom, oppositely to capacity. As a general trend, faster memories are smaller and more expensive, whereas slower ones allow larger capacity and are cheaper. Therefore, the technologies that are at the top of the pyramid are used as primary storage and contribute to increase computational performances, while the memories at the bottom are employed for long-term storage. Figure 3b shows a general architectural design, where different memory units are pipelined. The lower the layer index, the lower the latency and size, and the closer the memory is placed to the CPU. This way, the memory wall can be minimized, forming the illusion of an overall fast and large memory, where the speed equals that of the fastest memory, and the capacity that of the largest one.

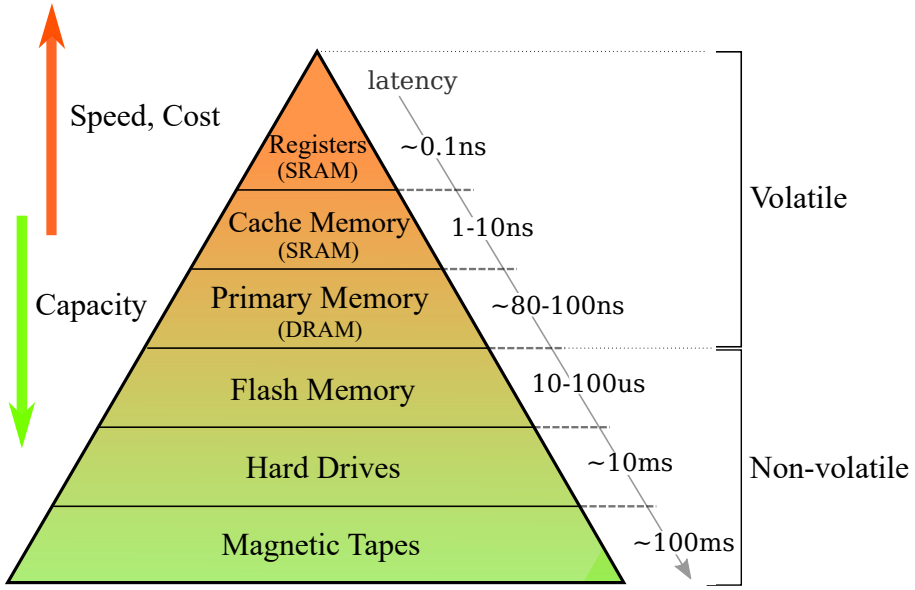
1.3 RANDOM ACCESS MEMORIES

Are called Random Access, those memories whose reading (writing) time does not depend on the physical position of the data stored (to store) [22, 26, 41]. This benefit derives from the layout of Random Access Memory (RAM) systems, which is schematically illustrated in Figure 4. The memory content is stored in a nxm matrix, where n is the number of rows and m the number of columns. At the intersection of each row and column, a memory *cell* stores either a bit or a word [4] of data.
The technical names for the routing rows and columns are, respectively, *word*, and *bit lines*. By selecting, i.e. opportunely biasing, the corresponding word and bit lines, one cell can be accessed at a time. In order to address the whole memory content with the minimum amount of bits, row and column decoders are usually implemented (illustrated in figure); specifically, n lines can be addressed with k bits, so that:

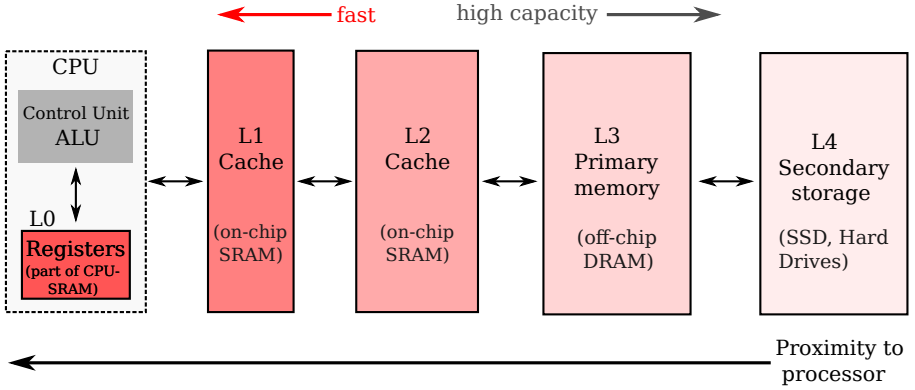
$$k = \log_2(n) \tag{1}$$

Once the voltage across the cell's bit and word lines is appropriate, its content can be read or written. Generally, a low-amplitude voltage is used to probe the cell state, while a relatively high amplitude is required for the write (or programming) operation. The non-selected lines are generally precharged at some fixed voltage, for example $Vdd/2$, so that no drop is guaranteed to appear across the inactive cells. The sense amplifier/driver amplifies the output swing to rail-to-rail values during a read process, while it buffers the write signal otherwise. The memory cell itself is usually in series with some access element, like a transistor or a diode, so that the leakage current from non-selected data is limited. Further details are given in Section 1.7, which illustrates the architecture of a random access memory array in the context of RRAM.

4. The term word indicates a series of bits, commonly one or two bytes (1 bite = 8 bits).



(a) Memory hierarchy, where speed and cost are found to trade with capacity. Faster memories, such as SRAM and DRAM, are generally volatile, while slower and bigger memories, like Flash, are non-volatile. Adapted and redrawn from [40]



(b) Schematic design of the memory organization in a modern digital computer. The memory hierarchy is exploited in order to reduce the memory wall problem, so that faster technologies, which are more frequently accessed, are placed closer to the CPU.

Figure 3 – The Memory Hierarchy

1.4 STORAGE CLASS MEMORIES

Looking at the memory hierarchy shown in Figure 3a, it can be noticed that there is a relatively large latency, as well as capacity, gap between the primary and the flash memory layer. In order to provide an additional tier for data storage, extensive research has been focusing on emerging memory technologies, whose performances would sit between DRAM and Flash [42–44].
In the literature, they are addressed with the term Non-Volatile Memory (NVM), in virtue of their non-volatility, or Storage Class Memory (SCM) [42–45]. Figure 5a shows how SCM locate in terms of speed, size and cost, with respect to SOA technologies. With storage capability similar to Flash, and access times approaching those of DRAM, at a fraction of its cost, SCM appears as a very promising class

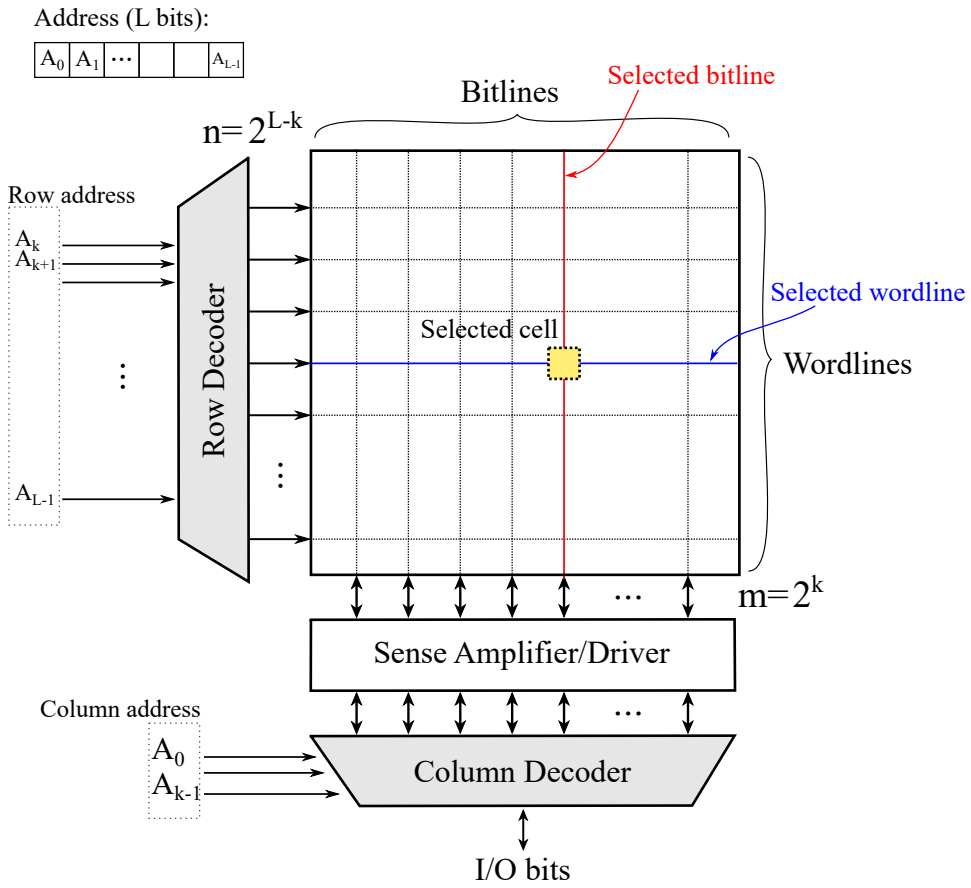
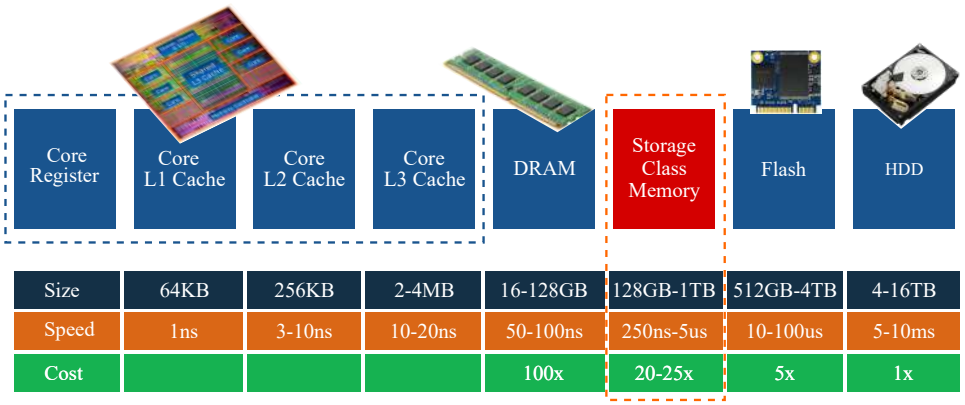


Figure 4 – A general design of a RAM memory array, where row and column decoders process the address bits, in order to grant access to the selected cell. A L-bit-long address is shown: k bits, from A_0 to A_{k-1}, are dedicated to addressing the bit line, while $L-k$ bits, from A_k to A_{L-1}, to the word line.

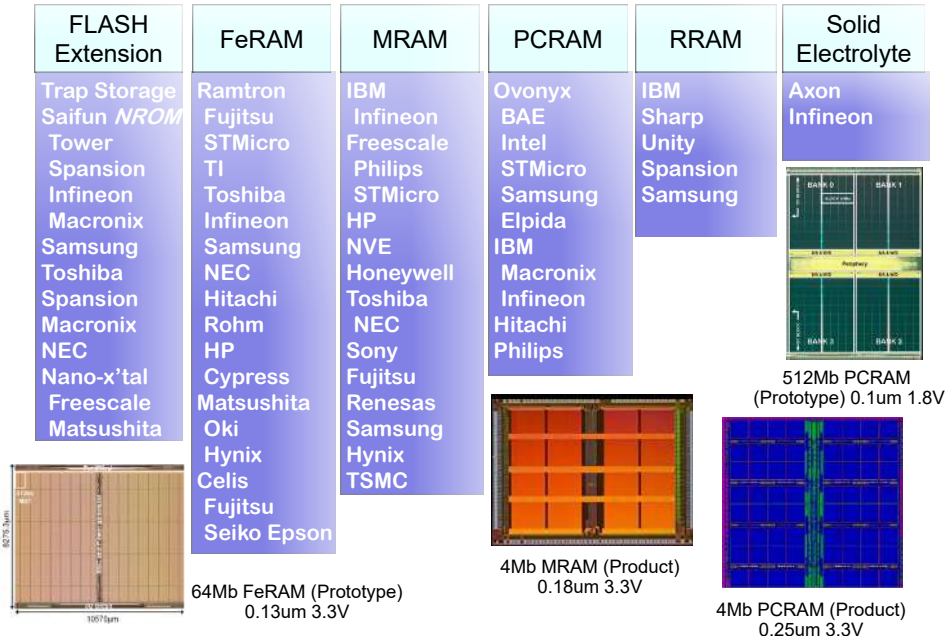
of memories. Indeed, SCM is expected to revolutionize the field, blurring the distinction between memory and storage tasks [40, 42–45]. A multitude of emerging technologies have been intensively researched, both in academia and by the industry. The most widely known are presented in Figure 5b: it can be seen that many semiconductor companies are active in SCM memory development [43].
The physical mechanisms that allow data storage widely vary depending on technology; in the following subsections, an overview is presented for FeRAM, MRAM and PCRAM. RRAM, being the focus of this thesis work, is discussed in larger detail in Section 1.5.

1.4.1 *Ferroelectric RAM*

Ferroelectric Random Access Memory, or FeRAM in short, stores binary information through the polarization of a thin layer by means of an electric field [46–51]. The cell structure is illustrated in Figure 6a, and resembles DRAM's, where a series transistor is used to gain access to a (ferroelectric) capacitor storing the logic state. Among the metallic plates, a thin ferroelectric layer is located; lead zirconate titanate (PZT) is a popular material choice [43, 44], although most recently



(a) Full memory hierarchy. Storage Class Memories (SCM) fill the gap between DRAM and Flash technologies in term of size, speed and cost. Adapted from [40, 45].

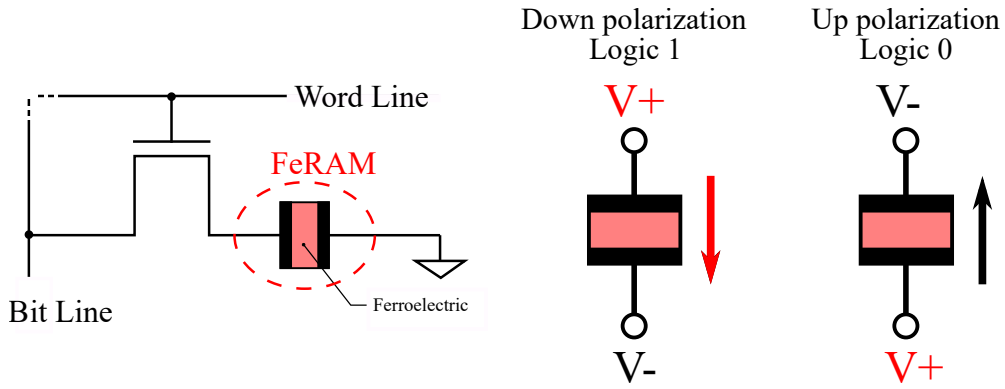


(b) List of main emerging technologies on nowadays market. Under each memory type, the relative major researching companies [43].

Figure 5 – Storage Class Memories: technologies and presence on the market.

hafnium oxide (HfO_2) has also been considered, thanks to its superior CMOS process compatibility and downscaling potential [51, 52]. By opportunely biasing the capacitor's electrodes, the resulting electric field changes the orientation of the atoms constituting the ferroelectric material, either to "up" or "down" direction, as illustrated in Figure 6b. When a change of orientation occurs, additional charge is induced, thus increasing the device capacitance. Distinct logic states can be associated to each orientation, and a change in polarization, or a lack of one, can be used to read the device state. For example, by applying a write voltage equal to the value required to write a logic 1. If the stored state does not match the polarity of the probing voltage, the polarization is flipped, and a current pulse can be detected as a result over the output line. Its presence, or absence, reveals what the cell state

was, although might require rewriting its content as consequence. FeRAM benefits from fast read/write times (~ 50ns) [49], low power consumption, and high number of write/erase cycles (~ 10^{14}), while major downsides are limited scalability and a destructive read process [43, 44, 46, 48, 50].



(a) A FeRAM cell, constituted by a ferroelectric capacitor, in series with an access transistor.

(b) Ferroelectric capacitor in opposite polarization state, which translate in binary logic values.

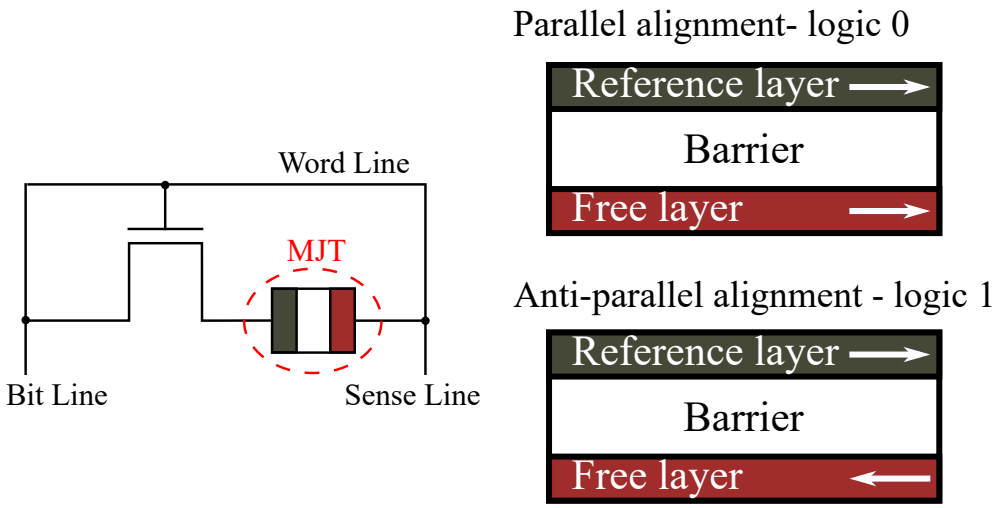
Figure 6 – FeRAM memory.

1.4.2 *Magnetic RAM*

Magnetic Random Access Memory (MRAM), also known as Spin-Transfer Torque RAM (STT-RAM), stores logic states though the orientation of magnetic domains in a thin layer, modulated by means of a spin-polarized current [43, 53]. A MRAM cell is shown in Figure 7: it implements a Magnetic Tunnel Junction (MTJ), which comprises of two ferromagnetic layers separated by a barrier layer. One of the magnetic layers, called reference layer, acts as a permanent magnet showing fixed magnetic direction, while the other's, called free layer, can toggle between parallel or anti-parallel. Depending on the orientation of the free layer, the cell's resistance changes: it is low when they are aligned (parallel direction) and high otherwise (anti-parallel). The cell content can be read by sourcing a non-desruptive DC voltage to measure the cell's resistance [54]. MRAM benefits from low power consumptions, high programming speed, and very high endurance as there is no known wear out mechanism [43, 54]. However, the required write currents can be quite high and thus possibly cause electromigration, especially for more advanced nodes [43, 55].

1.4.3 *PCRAM*

Phase Change Random Access Memory (PCRAM) relies on the state transition of a chalcogenide glass to encode logic values [43, 56]. The cell structure consists on a stack of at least three layers, where the chalcogenide material (a popular choice is $Ge_2Sb_2Te_5$ [57]) is sandwitched between two access electrodes. Figure 8a shows a typical cell configuration, featuring an access transistor. By heating the glass, through the flow of current over a dedicated heater, a small portion

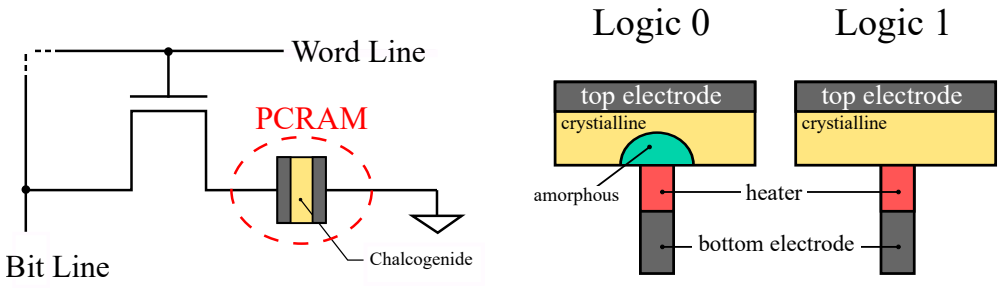


(a) A MRAM cell, constituted by a Magnetic Tunnel Junction (MTJ), in series with an access transistor.

(b) A MTJ in parallel and anti-parallel orientation, encoding different logic states.

Figure 7 – MRAM memory.

can be turned into either an amorphous or crystalline phase (Figure 8b). When the material is in the latter, its electrical resistance is low, as opposed to the amorphous phase; consequently, the two states can be associated to different logic values. The amorphous and crystalline resistances can be several orders of magnitude apart, which is a considerable advantage for the implementation in large arrays, and/or multibit storage. However, PCRAM usually suffers from *resistance drift*, where the amorphous phase resistance tends to decline over time, narrowing the read margin [58]. By sourcing a non-destructive DC voltage while measuring the resulting current, the cell state can be derived. PCRAM offers can endure a high number of write/erase cycles ($\sim 10^{11}$), while the main drawbacks are due to the high amount of current, in the mA range, required for the reset process, to physically melt the chalchogenide. In fact, this implicates relatively high power consumptions and possibly challenging downscaling due to high current densities [43, 56, 57].



(a) A PCRAM cell, in series with an access transistor.

(b) Binary states of a PCRAM cell: A logic 0 is associated to the amorphous, while a logic 1 to the crystalline phase.

Figure 8 – PCRAM memory.

1.5 RESISTIVE RANDOM ACCESS MEMORY

RRAM is an emerging, non-volatile memory technology where a logic state is encoded through a resistance value [59–61]: namely, a High Resistive State (HRS) and a Low Resistive State (LRS), where $R_{HRS} \gg R_{LRS}$, R_{HRS} being the cell resistance in HRS and R_{LRS} the resistance in LRS. A Logic 0 is usually associated to the HRS, also called *reset state*, whereas a logic 1 to the LRS, or *set state*. A resistance change can be triggered by opportunely biasing the cell's electrodes: a HRS to LRS transition (i.e. a set process) occurs by posing a set voltage (V_{set}) across the cell terminals, while a LRS to HRS (i.e. a reset process) is accomplished by sourcing a reset voltage (V_{reset}). A basic representation is illustrated in Figure 9: by supplying the appropriate bias, a thin conductive filament can be formed or dissolved inside the *switching layer*, thus modulating its resistance [59, 62, 63]. Figure 10a shows a Transmission Electron Microscopy (TEM) picture of a RRAM cell in LRS, where a thin filament can be seen bridging the access electrodes.

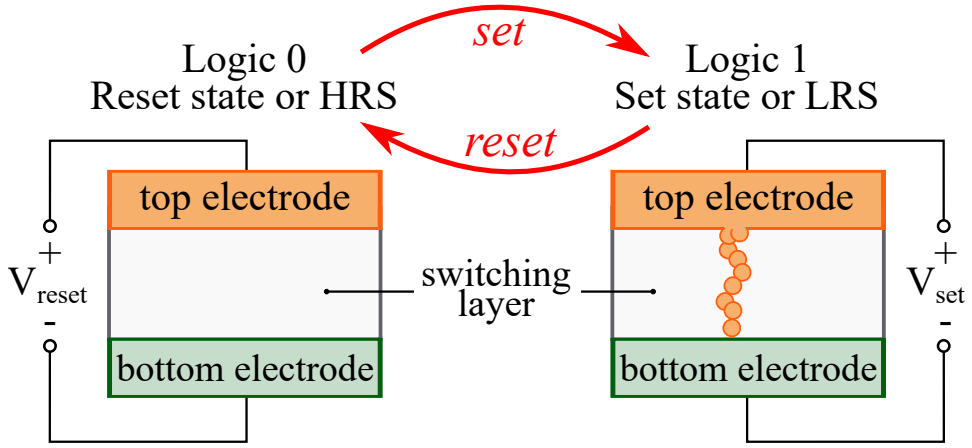
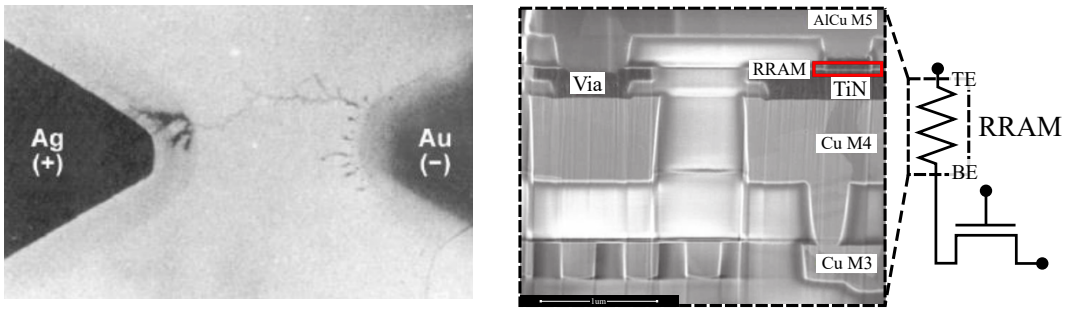


Figure 9 – Basic structure of a RRAM cell. A thin conductive filament can be formed or dissolved inside the switching layer, representing different logic states.

The phenomenon behind the operation of RRAM is known in the literature as *Resistive Switching* effect. Since the late 60s, when it was first reported that oxides can become electrically conductive under the effect of an electric field [62–64], the Resistive Switching effect has been recognized in a large variety of materials. RRAM technology has thus actually been around for several years, despite its role as next-generation memory being relatively recent; technological advantage and material compatibility have allowed the co-integration with the standard CMOS fabrication process, thus making RRAM an attractive technology for today's market [65]. Figure 10b shows a Scanning Electron Microscopy (SEM) picture of a RRAM cell, integrated on top of an underlying access transistor in BEOL.
Nowadays, RRAM represents a competitive alternative in the emerging memory market, exhibiting high programming speed ($\sim 10ns$), fair endurance ($\sim 10^6$ write/erase cycles), low power consumption ($< 10pJ$) and low cost.



(a) TEM picture of a RRAM cell in LRS, showing a conducting filament connecting the electrodes. [66]

(b) SEM picture of a RRAM cell, integrated in BEOL on top of an underlying transistor (130nm), between M4 and M5. Adapted from [67].

Figure 10

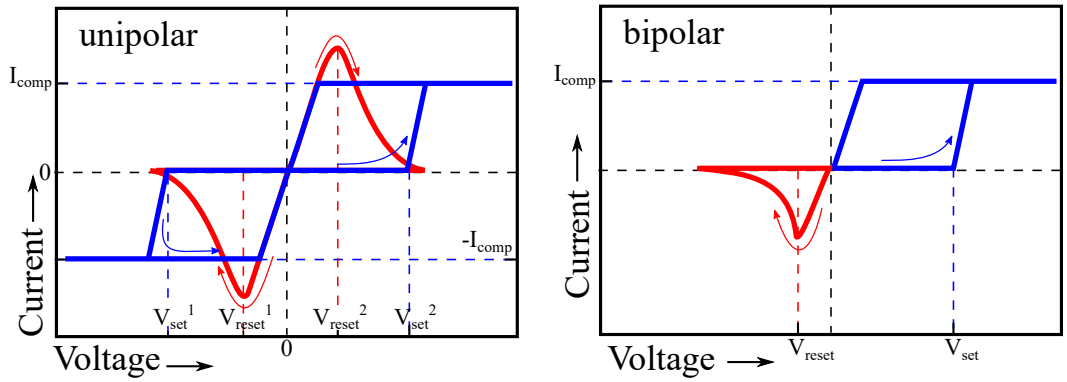
1.5.1 *Types of resistive switching*

Physically, a HRS to LRS transition is attributed to some sort of soft breakdown process [60, 61, 68] which can thus be made reversible, albeit inescapably damaging the cell over write/erase cycling, and finally limiting its endurance. The first set event a virgin[5] RRAM cell experiences is called *forming* or *electroforming process*, and it requires a voltage amplitude, the *forming voltage*, higher than following set voltages [60, 61, 68].
Two main types of switching can be distinguished, depending on the polarity of the set and reset voltages: unipolar and bipolar. The unipolar case is illustrated in Figure 11a: in this situation, the write/erase voltages have the same sign, although the amplitude of the reset voltage is usually lower than the set. The reasoning behind is that Joule heating is identified as the predominant mechanism behind the reset process [68, 69]. The bipolar case is shown in Figure 14b: here, set and reset voltages have opposite sign, due to the fact that the reset operation is attributed to the migration of charged species, although Joule heating can assist the process [68]. Our study focuses on bipolar switching RRAM, namely OxRAM and CBRAM, whose working principle is exposed in further details in the subsequent paragraphs.

OXRAM also kown as Valence Change Memory (VCM), OxRAM is a class of RRAM that exhibits bipolar resistive switching, whose filament is constituted by oxygen vacancies[6] [60, 61, 66, 68–70]. An OxRAM cell is shown in Figure 12, at point 1: the bottom electrode is made of an electrochemically inert, like Pt or TiN, while the top electrode of an electrochemically active material, such as Ta, Ti, Cu or Hf. The latter allows the formation or redox reactions and ion exchange with the switching layer, which is constituted by some transition metal-oxide, like HfO_x, TaO_x or AlO_x [66, 69, 70]. A fresh sample is highly resistive, showing a resistance value (IRS resistance) which is usually larger than any subsequent HRS.

5. A fresh RRAM cell displays a high resistance, generally much higher than the HRS that achieved after a reset operation. This state is sometimes referred to as Initial Resistive State (IRS).

6. A vacancy is a point defect in a crystalline solid, which occurs when an atom is missing from its original lattice site.



(a) Unipolar switching, where the set and reset voltages are either both positive or both negative.

(b) Bipolar switching, where the set voltage has positive sign while the reset negative.

Figure 11 – Unipolar and bipolar resistive switching.

Oxygen vacancies can either be already present inside the switching layer, or later introduced by the forming process [68, 70].
Phase 2 illustrates the forming operation: the appliance of a positive voltage between the electrodes, V_{form}, results in an electric field inside the switching layer, which ionises some oxygen atoms and pushes them outside the lattice, leaving behind some positively charged oxygen vacancies [69, 70]. The negative charges (oxygen ions) are attracted to the top electrode, where they recombine as oxygen atoms. The vacancies create a conductive path which bridges the electrodes, so that the cell's resistance transits to its LRS (step 3). Phase 4 shows the the reset process, which is triggered under reversed bias. V_{reset} is a negative voltage that causes the oxygen atoms to ionize back, and recombine with the vacancies constituting the filament. This process interrupts the conductive path, bringing the cell's resistance back to an high value, the HRS. Any subsequent programming operation follows steps 2-5, albeit requiring a lower voltage, $V_{set} < V_{form}$, to bring the cell to its LRS.

CBRAM Conductive Bridge Random Access Memory (CBRAM) is a bipolar switching class of RRAM, where the filament is build as a result of nanoscale electrochemical reactions [60, 61, 68]. The cell structure resembles a tiny ionic battery [71], where the electrodes are separated by a switching layer which conducts ions from one metal plate to the other, and is thus also called (solid) electrolyte. Figure 13 shows the programming flow of a CBRAM memory cell, reporting on the x-axis the voltage drop from the top to the bottom electrode, $V_{te} - V_{be}$, and on the y-axis the current, V/R_{cell}, where R_{cell} is the cell resistance. The top electrode is made of some electrochemically active[7] metal Me, for example Cu or Ag, while the bottom is electrochemically inert, like Pt or W. At point 1 of Figure 13, the cell is its Initial Resistive State (IRS), which is highly electrically insulating[8]. Once the voltage across the cell is risen towards positive values, the oxidation of the top electrode

7. With electrochemically active, is intended a material which can undergo reduction-oxidation (redox) reactions, usually triggered by the appliance of an external potential. Contrarily, the material is called electrochemically intert.

8. Prior to any set event, the range can extend to tens of GΩ [60].

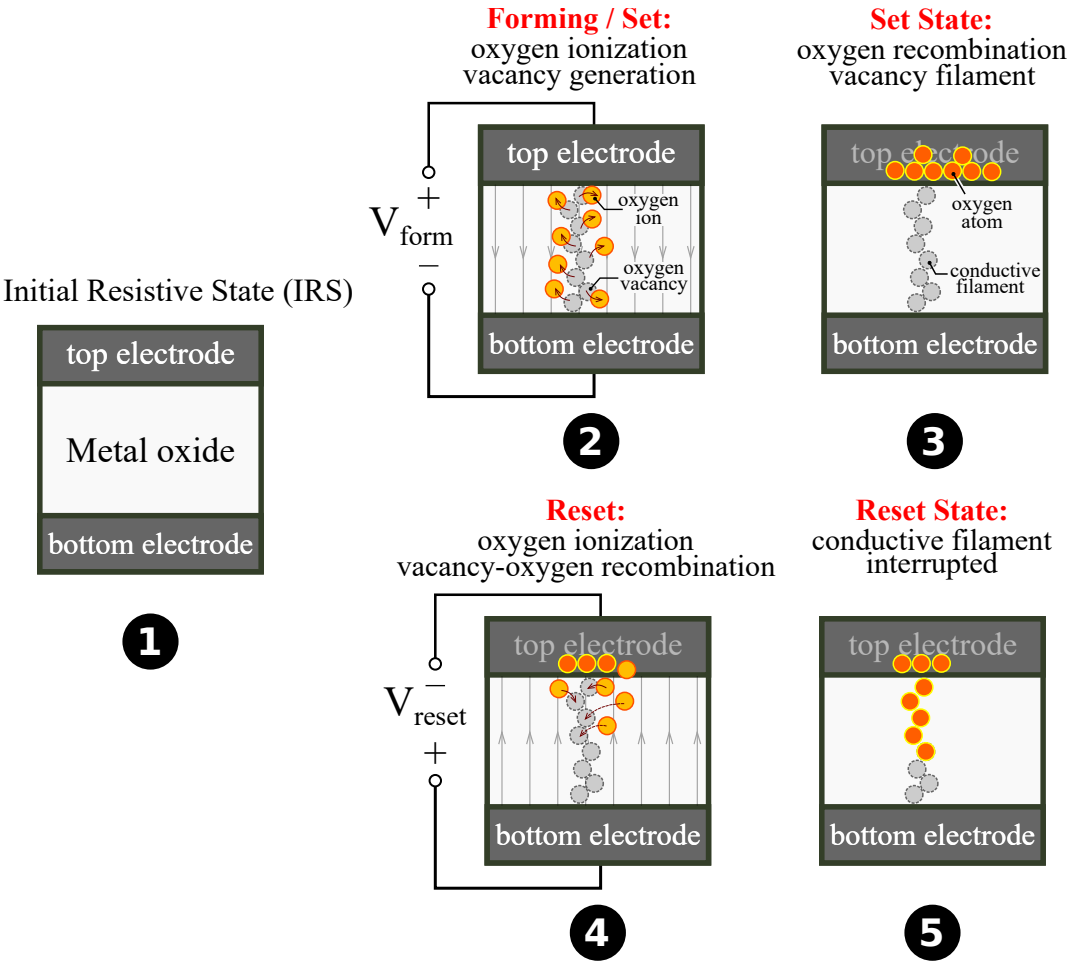


Figure 12 – Illustration of the operation of a bipolar OxRAM memory cell.

is triggered (point 2), and some metal atoms constituting the top electrode are ionized into Me^{z+} cations. In parallel, reduction occurs at the bottom electrode, so that the Me^{z+} ions lose their charge and revert to metal atoms.
By effect of the electric field inside the electrolyte, the metal ions pile up in filamentary fashion, and eventually short-circuit the cell terminals. Point 3 illustrates the set event, which occurs at voltage V_{set} and $R_{cell} = R_{LRS}$. The current saturates at compliance, I_{cc}, by effect of some external limiter, like a transistor or a resistor, connected in series with the cell in order to prevent its damage. As the voltage value is hereby varied in order to keep constant the value of the current, the line is dashed. The cell remains in LRS until point 4, where a negative voltage triggers the specular reactions of point 2, so that the filament atoms are ionized away and interrupt the conductive path, rising the cell resistance R_{cell} to R_{HRS}. After the electroforming event, some remaining of former filaments generally persist inside the switching layer, causing the HRS resistance to be lower than the IRS of a virgin cell, howbeit resulting much greater than the LRS, until the cell failure eventually occurs.

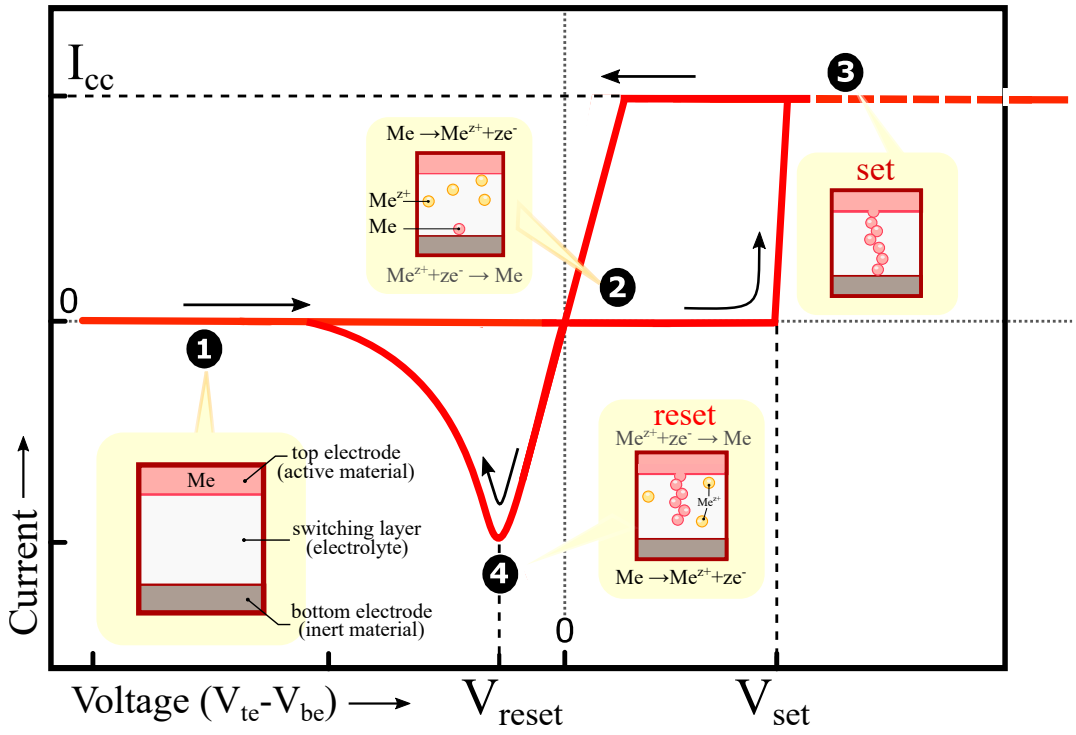


Figure 13 – I-V plot of a CBRAM memory cell, schematically illustrating the programming flow.

1.6 BENCHMARK METRICS

This section defines fundamental parameters used to characterize and benchmark memory technologies, in the context of RRAM.

- **Cell size**: refers to the amount of silicon surface occupied by a memory cell. The lower the cell size, the denser and more area efficient the memory array. The area is usually expressed by means of the technological *feature size* F, which corresponds to half the pitch of the lowest order metal line[9]. A RRAM cell can potentially be fitted into minimally size metal cross-points, so that its area can be as low as $4F^2$. Section 1.7 gives representation and further explanation on the design of RRAM arrays.
- **Window Margin (WM)**: defined as R_{HRS}/R_{LRS}, expresses the (resistance) gap between the binary states. In order to evaluate the robustness against technological dispersion, the WM is usually evaluated at the critical tail of the distribution; for example: $WM(+2\sigma) = R_{HRS}(-2\sigma)/R_{LRS}(+2\sigma)$. The higher the WM, the more robust the technology against reading errors, facilitating the integration in large arrays.
- **Endurance**: expresses how many times a cell/array[10] can be successfully programmed, before failure occurs due to accumulated damage brought by

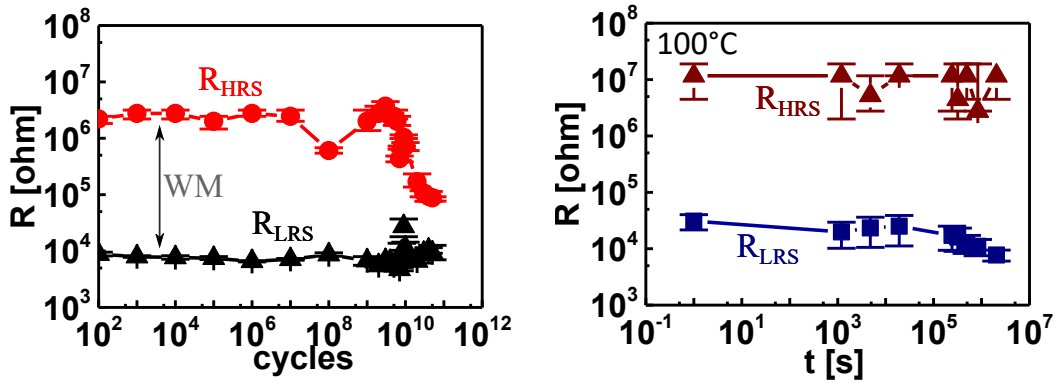
9. Metal 1 is the lowest order metal line for a given technology. The minimum pitch, 2F, is defined by the layout design rules.

10. When multiple cells are considered, the endurance is expressed along with its statistical dispersion.

previous write/erase processes. In the context of RRAM, it corresponds to the maximum number of set-reset cycles that a memory can withstand, before the HRS collapses onto the LRS (a threshold can be set, for example, at $WM(+2\sigma) = 10$). A well-known failure mechanism is a high current density inside the filament, typically taking place during a set event, which irreversibly damages the cell [68]. To extend RRAM lifetime, a fast-responding current clamp, like a resistor or a transistor, is introduced in series with the cell.

- **Retention**: is the amount of time a memory cell/array maintains the written logic state, before a spontaneous information loss occurs. For non-volatile technologies, a typical target time is 10 years. In the case of RRAM, the problematic state is constituted by the LRS: since, by design, a low amount of atoms are responsible for a dramatic resistance change, limited atomic rearrangement can cause the conductive filament to spontaneously dissolve over time [68]. Retention can be measured by performing a read operation at high temperature (for example 85°C) at regular time intervals, and then projecting the time at which failure occurs to a several-years timescale [68].

Figure 14 shows experimental endurance (14a) and retention (14b) characteristics for RRAM, where the window margin closure limits the technology performances.



(a) Endurance characteristic for a RRAM array, where the Window Margin (WM) reduces over cycling.

(b) Retention characteristic for 10 cells, where the WM slightly reduces over time.

Figure 14 – Endurance and retention characteristics for RRAM, adapted from [72].

In order to give a readable comparison chart, Table 1 reports performance metrics of various State Of the Art (SOA) emerging technologies. [68, 73].

1.6.1 *Variability and performance Trade-offs*

Ideally, the window margin, endurance and retention should be maximised. However, reported evidence has shown that these parameters are found to trade with each other; for example, a higher WM can be obtained by performing a set event at a higher current compliance, while endurance is degraded [72]. Nonetheless, above some threshold, both endurance and WM result degraded (for example,

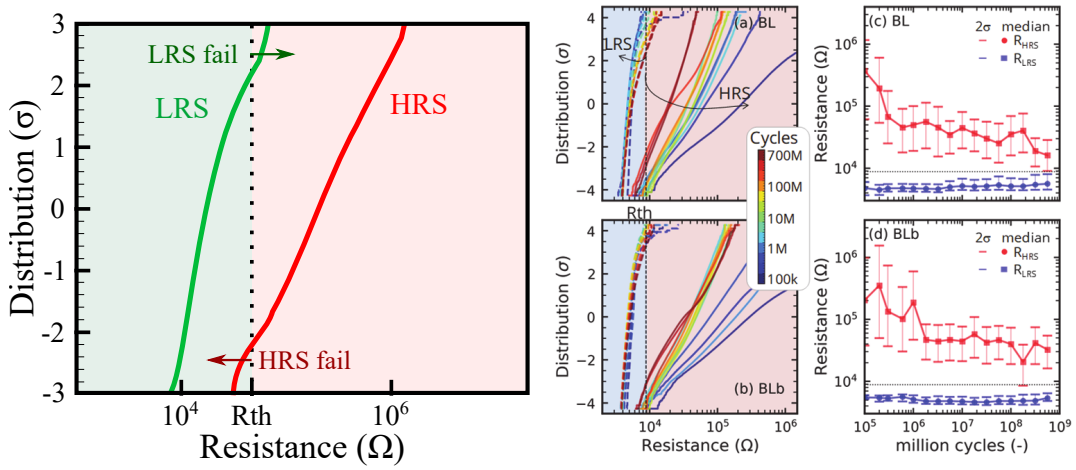
Technology	SRAM	DRAM	Flash NAND	Flash NOR	MRAM	FeRAM	PCRAM	RRAM
Non-volatility	no	no	yes	yes	yes	yes	yes	yes
Operating Voltage [V]	~ 1	~ 1	~ 10	~ 10	1 – 3	1 – 4	3 – 5	1 – 5
Cell size [F^2]	> 100	> 6	5	10	6 – 20	> 15	4 – 20	< 4
Write time [ns]	~ 1	~ 10	10^5-10^6	10^4-10^6	~ 1	< 50	~ 50	< 10
Read time [ns]	~ 1	1 – 10	10^4	~ 50	< 10	~ 50	< 10	< 10
Retention [time]	–	~ 60ms	> 10y	> 10y	> 10y	> 10y	> 10y	> 10y
Endurance [cycle n.]	> 10^{16}	> 10^{16}	> 10^4	> 10^5	> 10^{15}	10^{15}	10^{11}	> 10^6
Energy [pJ/bit]	~ 10^{-3}	~ 0.01	~ 0.01	100	~ 0.1	FeRAM	10-100	1 – 10

Table 1 – Benchmark of main memory technologies.

Nail *et al.* [72] reported 200µA to be the max advisable compliance value for their technology). Moreover, endurance and retention are found to trade with each other at a given window margin.
Another important parameter for RRAM is technological *variability*, intended as the lack of uniformity between LRS and HRS values. Figure 15 shows the problematic: 15a reports the LRS and HRS distribution for 2k samples, while 15b illustrates the evolution of the distribution over cycling. It can be seen that the resistance values vary between different devices as well as programming cycles. Variability is a well known, critical aspect for RRAM, constituting the major hindrance to the implementation of large memory arrays [74–76]. In fact, high statistical variation can result in the overlap of logic values, making the binary states indistinguishable. RRAM is known to be afflicted by both intrinsic and extrinsic varibility [74–78]; while the former results from the stochastic nature of the resistive switching mechanism itself [67, 76, 78], and should be minimized through material engineering, the extrinsic part is caused by erratic programming conditions [79, 80]. Therefore, careful circuit design aimed at reducing the non-uniformity of programming variables (e.g. the wiring voltage/current), has the potential to narrow the resistance distribution [80]. This work tackles the issue of RRAM variability by proposing novel programming circuitries, which are presented in Chapter **??** and Chapter **??**: respectively, they rely on a charged capacitor and a Digital to Analog Converter (DAC) to perform the set operation, in order to minimize the writing energy and writing current dispersion.

1.7 ARRAY STRUCTURES

RRAM arrays are arranged in a similar layout to that presented in Section 1.3 for a general RAM memory, where cells are located at the cross-points of horizontal and



(a) LRS and HRS device-to-device variability, which gives rise to fails at the overlapping of the distributions. Adapted from [67]

(b) Cycle-to-cycle variability. As the cycle number increases, the WM closes. [67].

Figure 15 – Variability in RRAM.

vertical metal lines, the word and bit lines respectively. Each cell is constituted by a RRAM device, storing one bit of information, in series with a selection element, whose impedance is very high when the memory device is not accessed, and low otherwise. The presence of an access element is necessary in RRAM arrays, in order to avoid the programming of non-selected cells, as well as leakage contributions to the output current during the read operation, which might cause readout errors. The phenomenon is known in the literature as *sneak paths*. Figure 16 illustrates the problematic during a read operation, where a two-terminal crossbar array features cells simply consisting of a RRAM device.

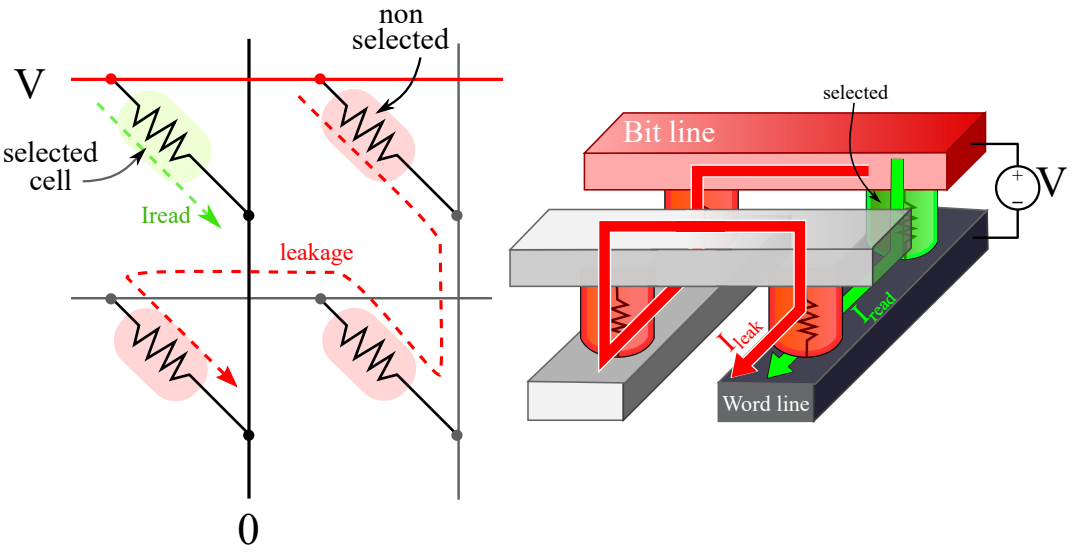
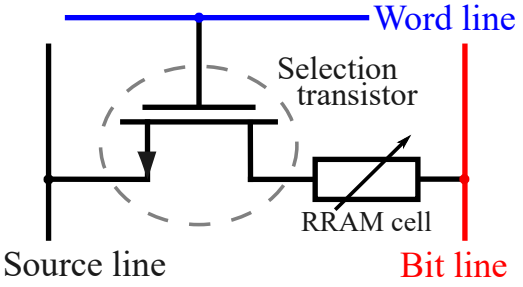


Figure 16 – Sneak paths problem in a selector-less array. Non-selected cells provide an alternative path for current, which results in a leakage contribution at the output.

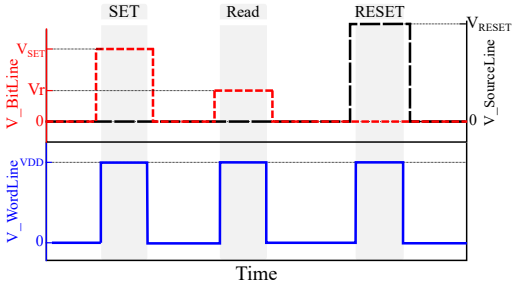
The selected cell is highlighted in green, and voltage V_r is posed across its terminals by biasing the corresponding bit and word lines. The output current flowing towards the word line is the sum of the ideal term, I_{read}, plus a undesired contribution I_{leak}, which arises by non-selected cells. As a result, the output current is higher than expected [11], possibly causing the readout circuit to mistakenly interpret the logic state of the selected cell.
The two main SOA cell configurations adopted in order to prevent sneak paths are illustrated in Subsection 1.7.1 and 1.7.2, where, respectively, the selection operation is carried out by a transistor and a volatile RRAM, called as *selector*.

1.7.1 *Transistor-RRAM: 1T1R cell configuration*

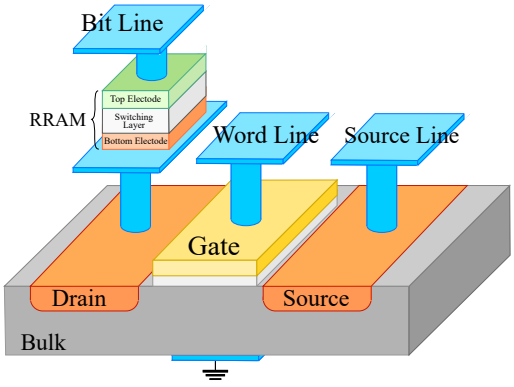
Figure 17a illustrates a popular memory cell topology, known in the literature as one-transistor-one-resistor (1T1R) configuration [81], featuring a MOSFET in series with a RRAM device. Figure 17b shows the corresponding biasing of the metal lines during the set, reset and read operations.



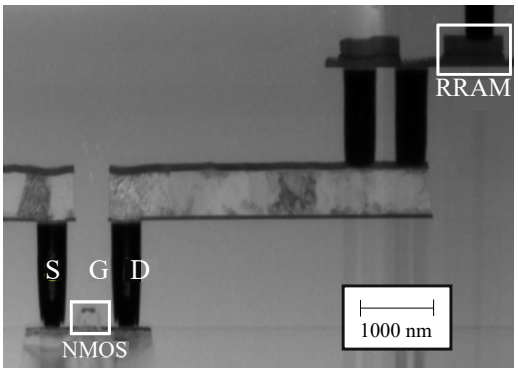
(a) Memory cell, comprising of a RRAM resistor in series with a NMOS transistor, which acts as selection device.



(b) Metal lines biasing during the set, read and reset operation of a 1T1R cell, where the transistor (N-MOS) is used as selector.



(c) 3D schematic view of the integration process of a 1T1R cell, where the RRAM element is stacked on the drain of the underlying NMOS transistor.



(d) TEM picture of 1T1R cell, where the RRAM is stacked on top of the transistor's drain. Adapted from [81].

Figure 17 – 1T1R cell configuration.

11. The lower the impedance of the non-selected cells, the higher the leakage contribution, thus making the presence of sneak paths more critical if the RRAM cells are in LRS.

When the word line voltage, $V_{WordLine}$, is high, a conductive channel bridges the drain and source of the access transistor, allowing the voltage difference between the bit and source lines, $V_{BL} - V_{SL}$, to drop across the RRAM. The souce line is grounded during the set and read operations, whereas it is positively biased during the reset [12]. This allows to flip the polarity across the RRAM element without implementing negative voltage sources, at the price of a relatively narrowed swing [13].
The RRAM is generally stacked on top of the transistor during a Back End Of Line (BEOL) process, as illustrated by Figure 17c. The area footprint of a 1T1R cell is $6F^2$: choosing a minimally sized NMOS allows to maximize the integration potential, although awareness to any significant voltage division between the impedance of the transistor and the memory itself is demanded. Figure 17d shows a Transmission Electron Microscopy (TEM) photo of a 1T1R cell, highlighting the front-end transistor and the back-end RRAM.

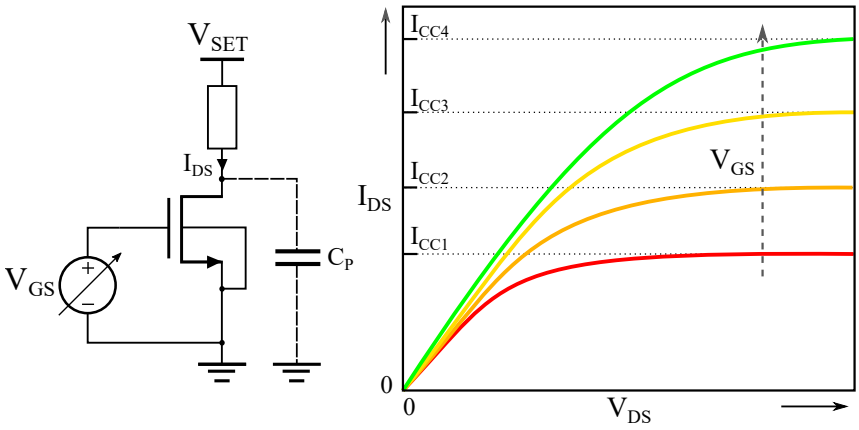


Figure 18 – Equivalent circuit of a 1T1R cell, where the voltage on the gate is varied to obtain different levels of current compliance. Capacitor C_P represents the parasitic capacitance on the drain, which limits the transistor's frequency response.

As illustrated by Figure 18, the access transistor can be used to clamp the current during the set operation. This is done by opportunely biasing the gate (V_{GS}), so that when the transistor enters in saturation, the drain current (I_{DS}) equals the desired compliance value. This approach has been largely adapted in the literature [66, 70, 81] due to its simplicity and relative effectiveness; in particular, an integrated transistor offers minimal parasitic capacitance C_P loading the drain, so that a reduced current overshoot can be obtained when the memory sets (further details on this aspect are given in Chapter **??**). However, this method is arguably far from ideal, as technological variability and testing conditions (like temperature) can potentially produce large statistical dispersion in the drain current, especially for smaller nodes. This problematic is further expanded in Chapter **??**, where a novel programming circuit is proposed, in order to overcome the dispersion of programming parameters.

12. It should be pointed out that, due to its structural symmetry, when the source line is positively biased with respect to the bit line, the drain and source of the transistor are swapped.

13. When a NMOS is implemented, the voltage amplitude at the transistor's source must be lower than the gate's by at least a threshold voltage (V_{th}) in order to keep the MOS on. Consequently, the amplitude swing is reduced to $< 0, V_{DD} - V_{th} >$

1.7.2 *Selector-RRAM: 1S1R cell configuration*

A *selector* is a two-terminal device, which can be put in series to a RRAM cell in order to suppress sneak paths [82, 83]. As a source line is not required, memory arrays can be arranged in *crossbar* topology, which is illustrated in Figure 19. This layout offers high scalability, as the word and bit lines can potentially be minimally spaced (pitch=2F), resulting in a cell area of $4F^2$ [83].

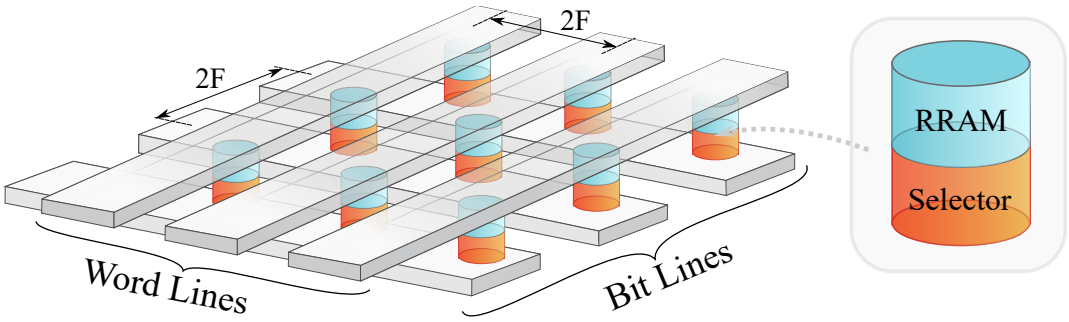
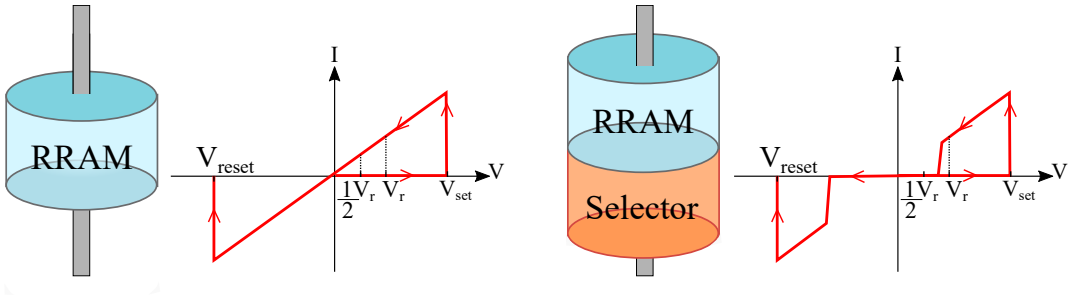


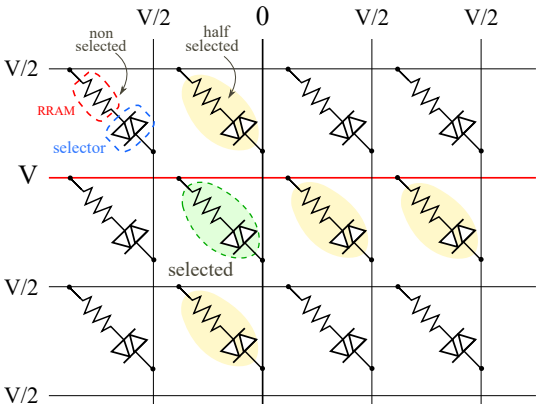
Figure 19 – Crossbar memory array, where each cell comprises of a RRAM device in series with a selector. If the metal line widths and the spacing between them are designed using minimum feature size (F) the unitary cell footprint results equal to $4F^2$.

The ideal selector resembles a (bidirectional [14]) diode: its resistance should be negligible once enough voltage drops across its terminals, and very high otherwise. It might be assumed that a silicon diode would be a suitable selector: unfortunately, this is not the case, as a p/n junction is produced during front end, while a RRAM in back-end-of the line. Consequently, optimal co-integration with the memory cell would be achievable only by means of some kind of "back-end diode". Different technologies have been emerging, the most promising being Ovonic Threshold Switching (OTS), whose overview is given in Subsection 1.7.3.
Figure 20a shows the IV curves of a selector-less RRAM (on the left) and a stacked selector-RRAM element (on the right). It can be seen that, when the selector is added, the current across the device remains low until the voltage hits threshold $\pm V_{th}$, at which point the selector becomes conductive, and the curve traces that of the single RRAM's. The selector shows low conductivity when the voltage is below V_{th}, so that the sneak-path effect can be mitigated, preventing programming disturbance and reading errors.
A good selector would offer a high R_{off}/R_{on} ratio, R_{off} being its resistance when in insulating state, while R_{on} the resistance in its conductive state.
The two most popular biasing schemes, namely the V/2 and V/3 polarization, are illustrated in Figure 20b and Figure 20c, respectively. In either case, the word line of the selected cell is set to write voltage V, while the bit line is at ground. The difference between the schemes lies in how the non-selected lines are biased: according to the V/2 case, any other access line is set to V/2. As a result, the number of half-selected cells, biased at voltage V/2, is $N_{hs} = n + m - 2$, while the rest, $N_{ns} = (n-1)(m-1)$ see no voltage drop.
On the other hand, in the V/3 scheme the non-selected word lines are biased at a

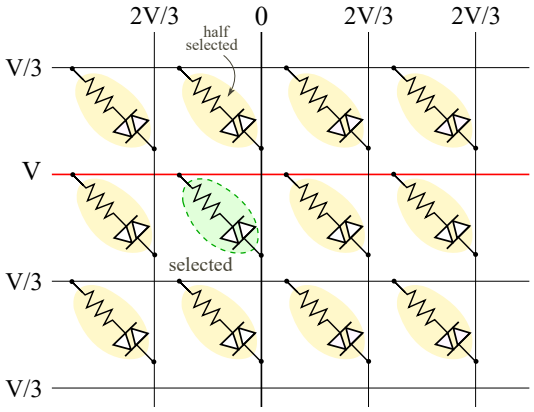
14. In the case of a bipolar RRAM.



(a) On the left: I-V characteristic of a RRAM element, on the right: I-V characteristic of the pair selector-RRAM.



(b) Crossbar array in V/2 polarization scheme.The half-selected cells see a voltage drop equal to half of the write voltage.



(c) Crossbar array in V/3 polarization scheme. With exception of the selected cell, the remaining cells are half-selected to a third of the write voltage.

Figure 20 – Operation of the RRAM-selector pair in a memory array.

third of the write voltage, while the non-selected bit lines at two thirds. As a result, more cells are partially selected: $N_{hs} = nm - 1$, although the unitary leakage from a non-selected cell (at a given RRAM array) is reduced with respect to the V/2 scheme, as the voltage drop is lessened from V/2 to V/3.
In practice, the optimal biasing scheme should be chosen given the specifics of the memory array at hand, in particular the array size and the device non-linearity [83].

1.7.3 *OTS Selectors*

One of the most promising and intensively researched selector technologies is represented by Ovonic Threshold Switching (OTS) materials, which are based on chalcogenide glasses [84–88].
A typical I-V characteristic is shown in Figure 21: the selector transits from a highly restive (or off) state, to a highly conductive (or on) state, once its voltage and current are greater, respectively, than thresholds V_{th} and I_{th}. It remains "on", until the current falls below the holding value I_h, at which point the resistance reverts

to a high value. The leakage current, I_{leak}, is expressed as the current that can be measured when the voltage across the OTS device is equal to $V_{th}/2$.

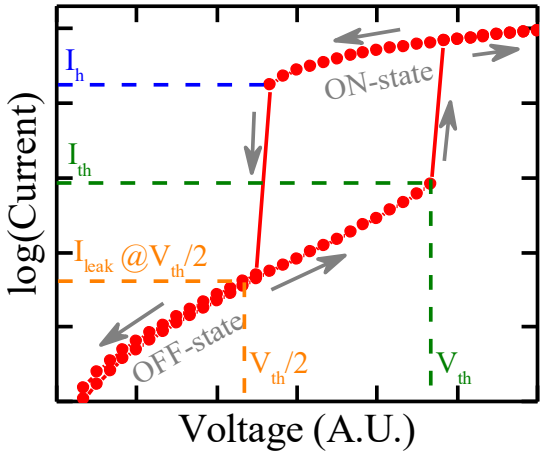


Figure 21 – typical I-V curve of a OTS selector [88]

1.8 OBJECTIVE OF THIS THESIS WORK

In this introductory chapter, the context of this thesis work was given, along with an overview on the main emerging memories in today's market. RRAM, the technology of choice in this study, was presented in larger details, along with some common architectures for the implementation of memory arrays.
As further discussed in Chapter **??**, the motivation behind our study on RRAM is partly rooted in its interest as Storage-Class-Memory, and partly in its potential as a novel type of energy source; the two behaviours are envisioned to coexist, as they share the same, faradaic-based, working mechanism. The attractiveness of a RRAM-based *nanobattery* lies in its high scalability potential, which could be exploited to obtain highly integrable, solid-state energy sources. Moreover, material compatibility with the standard CMOS process would offer the unique advantage of a dual memory/battery device, where the energy could be delivered in close proximity to the target, lowering the power dissipation and delay that occur over standard Power-Delivery-Networks.
In the subsequent chapters, the focus of this manuscript shifts, from device characterization, to integrated circuit design. Chapter **??** and Chapter **??** propose two experimental alternatives, namely a charged-capacitor and Digital-to-Analog Converter (DAC) based programming methods, for setting RRAM. The goal is to reduce both the wide variability which afflicts SOA technologies, and the energy waste that occurs during the set process.
The connecting thread between the diverse parts of this thesis is the research of advanced solutions for a new, energy-efficient, memory technology. As previously presented in this chapter, our quest answers a urgent need in today's semiconductor market, where power consumption and speed are found to trade with each other.
This work is structured in three parts: part i: *Introduction* (Chapter 1), part ii:

RRAM as Energy Source (Chapter 2), and part iii: *Energy-Efficient Design Solutions for RRAM* (Chapters **??** and **??**).

1.9 CHAPITRE 1 - RÉSUMÉ EN FRANÇAIS

Dans ce chapitre introductif, le contexte de ce travail de thèse a été donné.
Les mémoires emergentes non-volatiles devraient révolutionner le domaine des mémoires, offrant des performances qui se situent entre les DRAM et Flash, permettant un stockage de données plus important à une vitesse d'accès et d'écriture plus élevée que les mémoires non volatiles d'aujourd'hui. Après avoir présenté les technologies les plus recherchées (FeRAM, MRAM et PCRAM), les memoires resistives RRAM ont été discutées plus en détail. Les principales mesures de référence (area, Window Margin, Endurance et Retention) ont été présentées et expliquées, et différentes topologies de mémoire ont été comparées. de plus, des typologies d'architecture pour l'intégration de matrices de mémoire à base de RRAM ont été présentées, telles que les schémas 1T1R et 1S1R.

Comme présenté plus en détail au chapitre **??**, la motivation de notre étude sur le RRAM est en partie ancrée dans son intérêt en tant que Storage-Class-Memory, et en partie dans son potentiel en tant que nouveau type de source d'énergie ; les deux comportements sont censés coexister, car ils partagent le même mécanisme de fonctionnement basé sur le faradique que les batteries ioniques.
L'attrait d'une *nanobatterie* à base de RRAM réside dans son fort potentiel d'intégration, qui pourrait être exploité pour obtenir des sources d'énergie à l'état solide hautement intégrables. De plus, la compatibilité matérielle avec le processus CMOS standard offrirait l'avantage unique d'un dispositif à double mémoire/batterie, où l'énergie pourrait être délivrée à proximité de appareil cible, réduisant la dissipation de puissance et le retard qui se produisent sur les réseaux de distribution d'alimentation standard (Power Delivery Networks).

Dans les chapitres suivants, l'objectif de ce manuscrit passe de la caractérisation des dispositifs à la conception de circuits intégrés. Le chapitre **??** et le chapitre **??** proposent deux alternatives expérimentales, basées sur des méthodes de programmation basées sur un condensateur chargé et un convertisseur numérique-analogique (DAC), pour programmer des RRAM. L'objectif est de réduire à la fois la grande variabilité qui afflige les technologies SOA et le gaspillage d'énergie qui se produit au cours du processus de set.

Part II

RRAM AS ENERGY SOURCE

2

ENERGY STORAGE IN RRAM

This chapter illustrates the concept of a hybrid dual-behavior device, based on RRAM, for both data retention and energy storage. RRAM as energy storage element is a novel concept, which is grounded on the notion that its working mechanism as memory is based on faradaic reactions, similar to those taking place in ionic batteries. As further expanded by Sections 2.1 and Section 2.2, a RRAM-based battery would offer unique advantages, such as high scalability and CMOS process compatibility, which motivate our study on its feasibility.
Section 2.3 presents the techniques and instrumentation implemented for the evaluation of energy-storage capability, while Section 2.4 discusses results of various electrochemical characterizations, performed on SOA CBRAM samples. Although far from conventional solid-state battery framework, our results allow us to draw promising preliminary considerations, and express high potential for various standard and emerging applications.

2.1 THE INTEREST OF RRAM AS ENERGY SOURCE

Recent advances in fields like artificial intelligence (AI), the internet of things (IoT), virtual reality, and cloud computing have been demanding ever-increasing computational capability. Power-efficient calculation is key in any task implemented on a chip, from high-performance core computing to the edge of IoT. While process miniaturization managed to effectively reduce consumptions in the past [1, 2], its effectiveness has been dropping for recent technological nodes, the overall balance in power cost coming majorly from losses over transmission lines [3–5]. As a result, further downscaling is expected to soon no longer pay off the development investment, and alternative strategies must be considered. This branch of research involves the so-called Moore than More's devices [89].Some attempts to ease consumption rely on on-chip integration of nanoscale energy sources, [90, 91] while others, notably in memory computing (IMC) [9–11, 13], offer a revisited architecture where computation is decentralized. Other research groups proposed the integration of energy sources in close proximity of the memory,[15, 16] or completely integrated within the memory architecture [17].
‘ Our work attempts to tackle the need of energy-efficient computation with a radically different concept, where *dual-behavior* devices are able to store either information or energy, depending on the applied bias. Such capability would be greatly beneficial, allowing localized and high bandwidth energy supply to the processing unit (the memory or a dedicated arithmetic logic unit, ALU). The RRAM samples considered in our study indeed resemble ionic batteries at the nanoscale, providing ground to our inquiry of using these devices as energy sources, other than memory cells [71]. Their operation relies on faradaic processes; therefore, the resulting energy density is expected to well exceed that of electrostatic capacitors, possibly being comparable to supercapacitors [71, 92]. The diameter of the devices under

study can range between 1µm and 100µm, resulting much smaller than the diameter of SOA planar supercapacitors (mm^2) [71, 91, 92], making such architecture more scalable and granular than any other SOA integrated power source.
Energy storage is achievable when the device is, under a memory point of view, storing a logic 0, and could be accumulated during low logic operation activity for later use, for example, during the most power-hungry phases. These devices would also offer the advantage of placing the battery cell in close proximity to the target, meaning reduced IR drop and voltage undershoot, which develop in a typical inductive-impedance power delivery network (PDN) [93].
Finally, a broad range of applications could be envisioned, each demanding different energy requirements, with widespread specifications. The most suitable target field should be selected taking into account the output voltage, energy, and power delivered by such RRAM-based batteries. Figure 22 provides an outlook on some possible implementations, with some quantified ranges in terms of energy and instantaneous power requirements. The three main eligible domains being energy to memory [94], energy to logic [95], and neuromorphic computing [96–99].

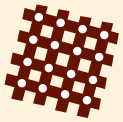
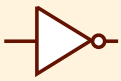
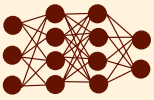
	Type	Energy [pJ]	Voltage [V]	Power [uW]
Energy to memory Cell switching energy 	NAND Flash	0.01	~20	1m
	NOR Flash	100	~5	100
	PCM	100	3-5	1k
	RRAM	1-10	2-5	100-1k
	STTRAM	0.1	1-3	10
Energy to logic Min. energy per operation 	ARM Cortex M0	3 - 6.4	0.4 - 0.7	0.64 - 240
	ARM Cortex M0+	1.1 - 11.7	0.19 - 1.2	0.4 - 735
	MSP430 compatible	7	0.32 - 0.48	56 - 497
	ARM Cortex M3	23	0.5 - 1	23 - 460
Neuromorphic Energy per synapic event 	SRAM-Based	27	28nm-compatible	63k
	SRAM-Based	105	14nm-compatible	-
	RRAM-Based	N/A	40nm-compatible	9.9k
	RRAM-Based	N/A	180nm-compatible	15.8k
	RRAM-Based	180	130nm-compatible	1.5k

Figure 22 – Estimation of required energy, operating voltage and instantaneous power for various application fields: memory [94], logic [95] and neuromorphic computing [96–99].

2.2 A NEW CONCEPT: IN-MEMORY ENERGY STORAGE

In Chapter i were presented a particular class of RRAM, whose working principle relies on electrochemical reactions: namely, OxRAM and CBRAM. Such devices appear to be the best candidates for energy storage purposes, as their operation as memeories resembles that of ionic batteries. In fact, the cell structure can be seen as a tiny battery, where the top and bottom metal layers constitute the access electrodes, and the (electrically insulating) switching layer is a solid-state electrolyte. Figure 23 reports a schematic illustration of the memory curve of a CBRAM, high-

lighting the electrochemical current contributions for low y values (zoomed view inset). In particular, two peaks are supposed to be revealed: one during the positive voltage scan rate (*oxidation peak*) and one during the negative (*reduction peak*), signalling the saturation of redox products inside the cell's switching layer.

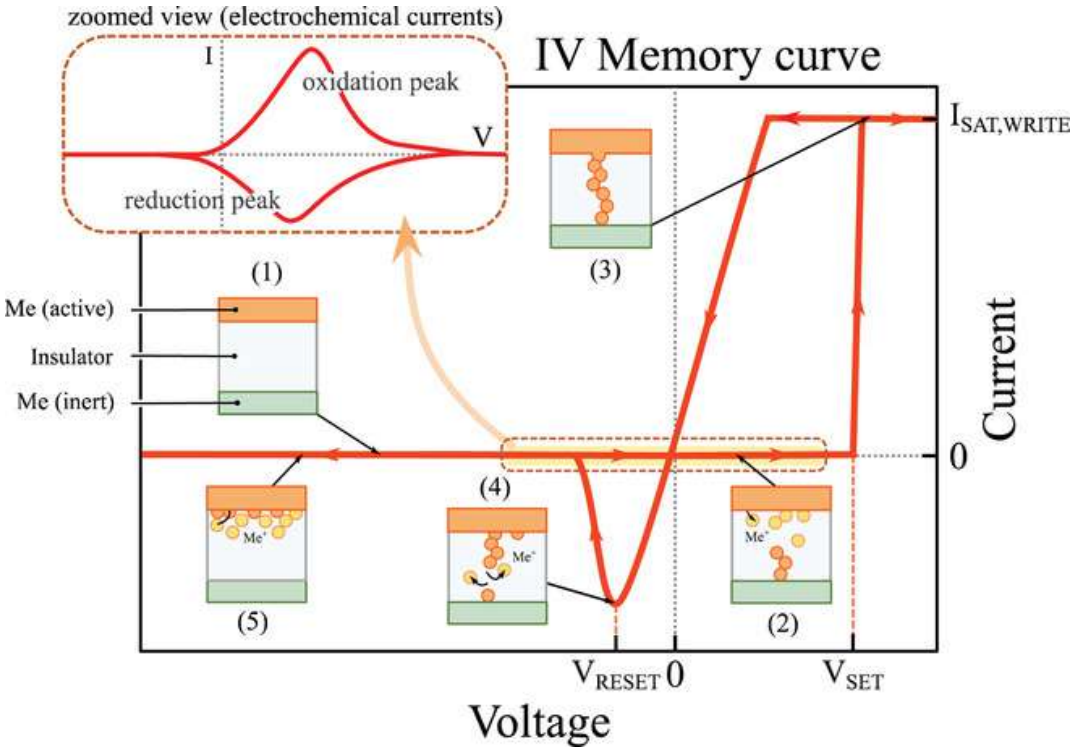


Figure 23 – A typical I–V curve of a RRAM. The memory is, initially (1), in a logic "0" state (highly resistive), and switches to a logic "1" in (3), when the filament bridges the electrodes. The memory goes back to "0" in (4), when the filament dissolves. In the inset, a zoomed view showing the electrochemical currents (redox peaks).

Figure 24 illustrates the envisioned device, where a crossbar-like array integrates elements which can either work as memory or energy storage cells. Sub-arrays are highlighted in green, for cells operating as battery, and in red for those operating as memory. The insets show a schematic of the different operating mechanisms: the *filamentary switching effect* being responsible for nonvolatile information storage, and the *Nernst* and *diffusion* potential for energy supply [60, 68, 100–102].
Valov et al. have reported measurements of open-circuit voltage across RRAM while the cell is in HRS and motivated their findings with the arise of Nernst, Diffusion, and Gibbs-Thomson potentials between the electrodes of the devices under study [100–102]. A diffusion potential resulted from a non-null gradient of charge inside the electrolyte, introduced by the release of redox ion products. Such potential can thus be externally controlled, following the activation of bias-triggered electrochemical reactions. As a result, a memory cell can be *charged* in order to release some energy at a later time. Once the potential across the electrodes has fallen below the activation energy, reactions cannot continue, and ions spontaneously diffuse back to a homogeneous concentration, gradually nulling the voltage over time. At the end of this transient, the cell enters the *discharged state*.

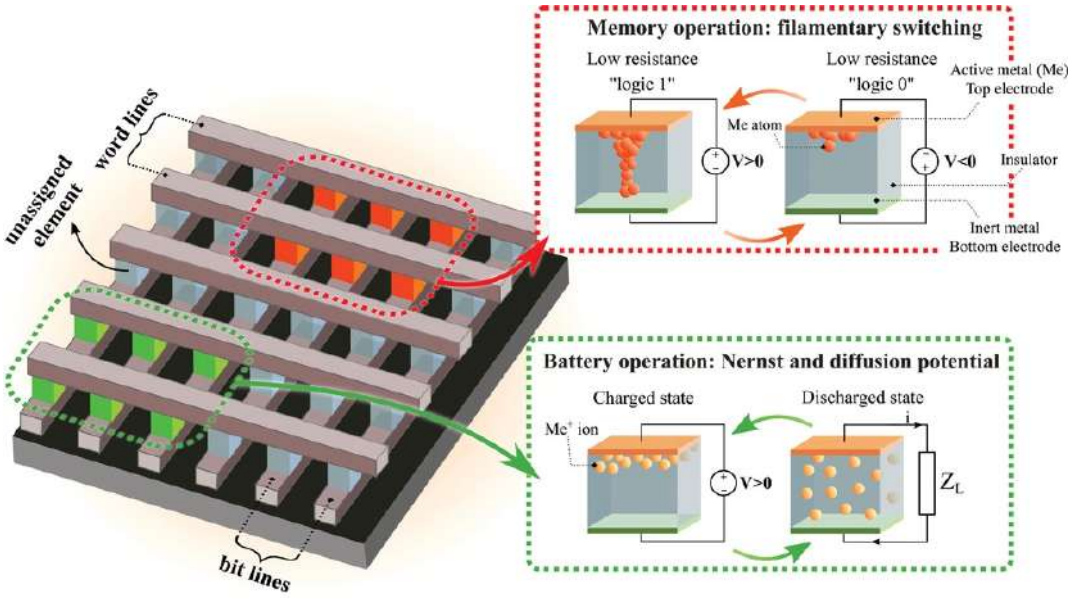


Figure 24 – Illustration of the innovative concept of a hybrid dual-behavior RRAM-based device. In the envisioned implementation, elements are placed in a crossbar array configuration, which allows maximizing energy and memory density. In red and green, respectively, cells storing logic values and energy.

Valov's group reported measurements of discharge currents, revealed when externally grounding the electrodes right after activating redox reactions, which they attributed to *re-equilibrium* movements of ions [100]. Although their research lies in a theoretical domain, meant to deepen the filamentary switching understanding, their findings inspired our study on the feasibility of a practical implementation of the nanobattery effect in RRAM.

2.3 EXPERIMENTAL DEVICES AND METHODS

We characterized SOA CBRAM devices, of stack: $CuTe_2Ge/Ta_2O_5/W$ [103–105]. Figure 42 illustrates a Transmission Electron Microscopy (TEM) picture of a sample, with Energy Dispersive Spectroscopy (ESD) elemental mapping. The dies were fabricated on 8-inch wafers, in 1R cell configuration, and three different Ta_2O_5 thickness (5, 10, and 15 nm) were produced and tested. Plug-shaped bottom electrodes (W) were first deposited. Consecutively, the Ta2O5 layer was RF sputtered, adjusting the sputtering time according to the targeted thickness. The density of the Ta_2O_5 layer resulted approximately equal to $7\ gcm^{-3}$.
A $30nm$ thick $CuTe_2Ge$ alloy was deposited as top electrode, and successively capped by a Ti/TiN layer. Seven different areas were produced, ranging from $0.07\mu m^2$ to $2.27\mu m^2$, by adjusting the bottom electrode diameter from $0.3\mu m$ to $1.7\mu m$, with a step of $0.2\mu m$.

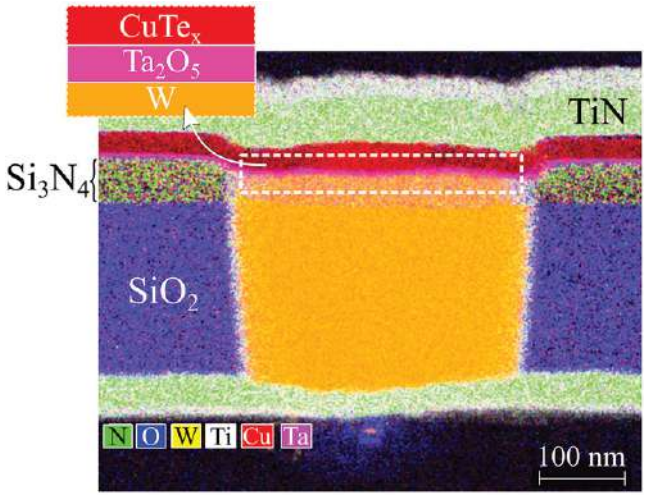
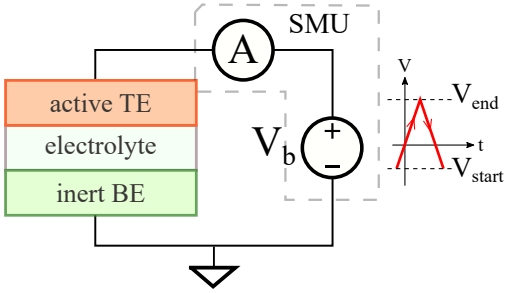


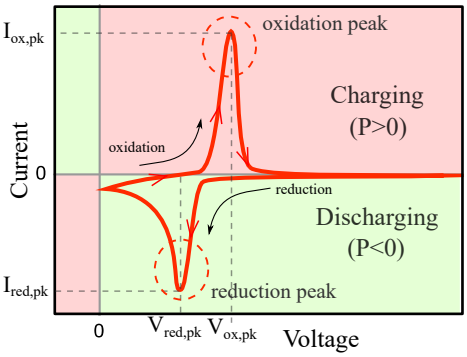
Figure 25 – TEM picture of a CBRAM sample, where material composition is highlighted through ESD mapping.

2.3.1 *Electrochemical characterization: Cyclic Voltammetry*

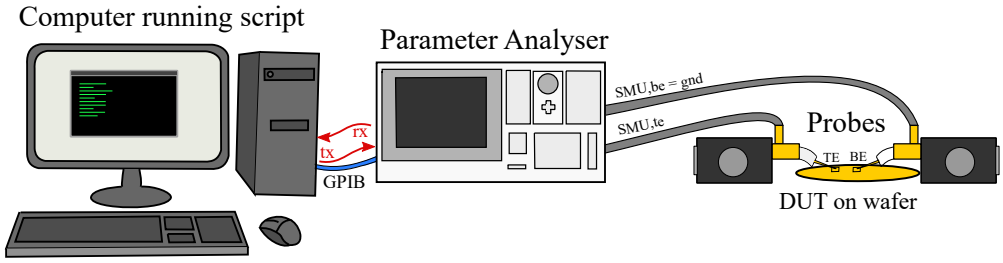
Cyclic Voltammetry (CV) is a widely used measurement technique in electrochemistry [106–109], which we adopted to evaluate both the battery and memory-like behavior for our samples. In electrochemistry, CV is to characterize oxidation and reduction processes in a solution [107, 108].
Figure 26a illustrates a schematic of the test setup for our samples: a triangular voltage, V_b, is applied to the chemically active electrode, while the (chemically) inactive is grounded; simultaneously, the current is monitored. Oppositely to usual CV measurements, a third reference electrode is missing, due to the solid-state nature of our electrolyte. The speed at which V_b is swept, also called *scan rate*, is kept constant over a test duration, and varied between different cycles/samples. Memory behaviour, i.e. a transition to the LRS, can be triggered if V_b reaches the sample's set (or forming, for a fresh cell) voltage.
Figure 26b shows a typical CV plot: when the potential is swept towards positive values, oxidation of the top electrode is triggered, and the (electrochemical) current increases, until it reaches an *oxidation peak* at $V = V_{ox,pk}$. The decrease in current from this potential on is caused by the lack of ions to reduce at the bottom electrode's interface, which delivers the counter-reaction needed to sustains the top electrode's oxidation. A dual situation happens when the sweeping direction in reversed: oxidation of the previously reduces ions takes place at the bottom electrode, while reduction happens at the top electrode. Another current peak is observed, at $V = V_{red,pk}$, and is thus called *reduction peak*. Since the power consumed during the top electrode oxidation is positive, this phase constitutes a charging process. On the contrary, as the power has negative sign during the (top electrode) reduction, we associate an energy release (discharge) to this phase. The double current peak shape shown in Figure 26b is typical of CV performed on batteries, and we expect to obtain a similar curve in for chemically promising RRAM samples.



(a) Schematic representation of CV setup for our samples. A Source Monitor Unit (SMU) was used, sourcing voltage V_b while reading the flowing current.



(b) Typical CV curve, showing oxidation and reduction peaks: the former is associated to a charge operation, while the latter a discharge (energy release).



(c) Full test setup.

Figure 26 – Cyclic Voltammetry tests setup for RRAM electrochemical analysis.

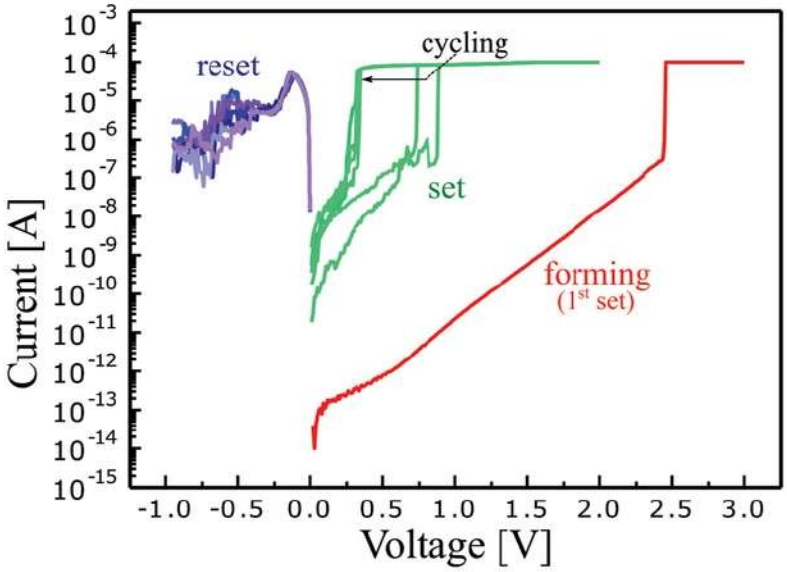
Figure 26c shows a more detailed representation of the test setup that was adopted to perform CV tests on our devices. Given the solid-state nature of our samples, the small physical scale and the high precision required, we implemented a Parameter Analyser's (HP4155) SMUs to perform the voltage sweep while reading the current flowing across the electrodes. This arrangement allowed to reach precision down to 10fA, while sweeping voltages in a range of $[10m - 1]V \cdot s^{-1}$. The wafers were loaded into a Cascade probe station, where micro-needles contact the device's top and bottom electrode access pads. Finally, the testing procedure was automatized by remote-controlling the parameter analyser with a computer.

2.4 EXPERIMENTAL RESULTS

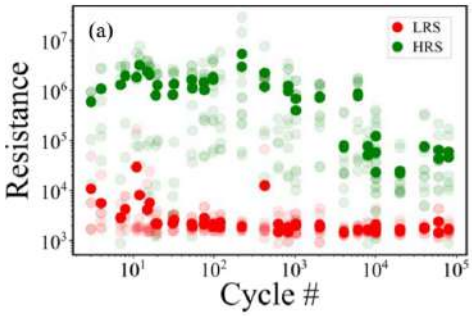
This Section presents the CV results collected for our Device Under Test (DUT)s, both as memory and energy storage elements.

2.4.1 *Operation as memory*

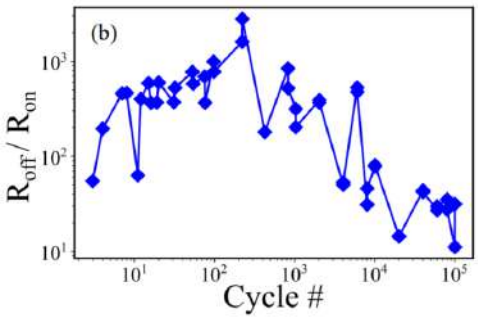
Prior to benchmark our technology as energy storage, we proceeded to validate memory operation. The setup shown in Figure 26a was adopted to gain insight onto the switching behaviour, and typical forming, set and reset voltages.



(a) Typical forming (in red), set (in green) and reset (in violet) curves of our Ta_2O_5-based CBRAM.



(b) Endurance test, showing HRS (green) and LRS (in red) resistances over 10^5 cycles. In darker color, the average values. The set voltage is equal to 2V, while the reset -2V. A programming pulse of 500ns was used for both operations.



(c) Median Window Margin (WM) over cycling, derived from Figure 27b. WM falls to 10 after 10^5 cycles.

Figure 27 – Endurance test results for the considered stack.

Figure 27a shows representative memory curves. The forming voltage results around 2.5V, while subsequent set occur below 1V, with the value decreasing over cycling. It follows that if a DUT is kept in pristine state (IRS), the CV voltage is allowed to reach 2V without triggering a LRS transition. Hence, pristine samples result more convenient when exploring a cell's electrochemical behaviour, as larger time and potential window could be analysed.
Figure 27b shows endurance tests, where 10^5 cycles are reported (in darker colors, mean values). Figure 27c shows a plot of the mean window margin, $WM_\sigma = R_{HRS}(\sigma)/R_{LRS}(\sigma)$, versus cycling. After 10^5 cycles, $WM_\sigma = 10$; which is a satisfactory memory performance.

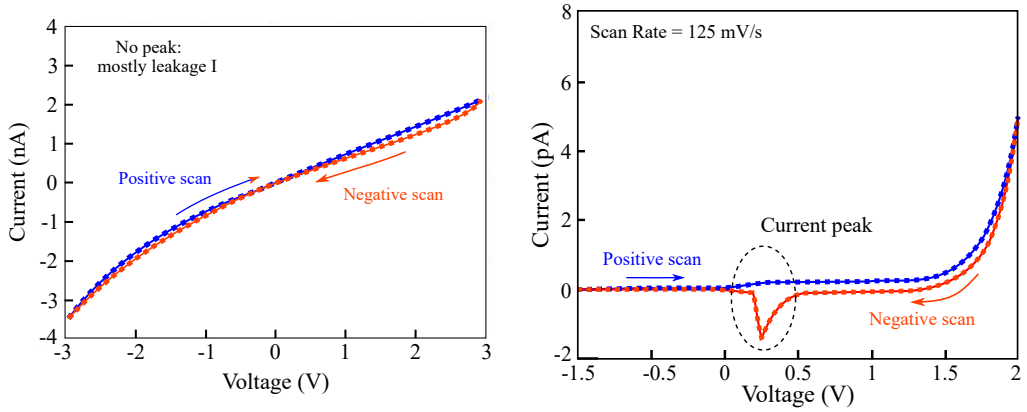
Device characteristics

Cell topology	Stack	Diameter [μm]	Thickness [nm]
1R	$CuTe_2Ge/Ta_2O_5/W$	$[0.3-1.7]$	$5, 10, 15$
1R	$Cu/HfO_2/TiN$	100	$5, 10, 15$
1R	$Cu/Al_2O_3/TiN$	100	$5, 10, 15$
1R	$Cu/SiO_2/Pt$	100	$5, 10, 15$
1R	$Ti/SiO_x/TiN$	$0.04, 0.12, 0.3$	5

Table 2 – Samples screened through CV tests: a wide range of materials and geometries were considered. Ta_2O_5-based samples later revealed as the most promising technology.

2.4.2 *Battery curves: reduction peak*

Once memory operation was confirmed, we proceeded with the electrochemical characterization of pristine CBRAM samples. Although the results reported in this thesis work focus on the Ta_2O_5-based stack presented in Section 2.3, our study began with the screening of a wide variety of SOA CBRAM, differing by material composition and geometrical dimensions. Table 2 summarises the characteristics of the devices analysed.



(a) Curve of a sample showing no elecrochemical currents: leakage and capacitive contributions are predominant, and no peak appears upon voltage sweeping.

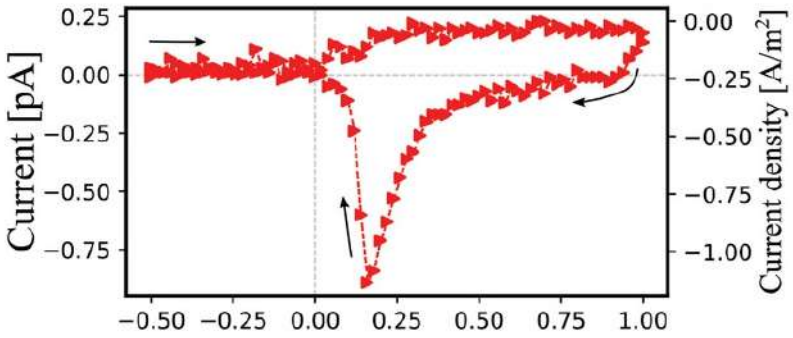
(b) Curve displaying promising electrochemical behaviour, typical of Ta_2O_5-based samples. A current peak appears upon negative sweeping, which cannot be assigned to leakage or capacitive currents.

Figure 28 – Typologies of IV plots obtained through CV tests.

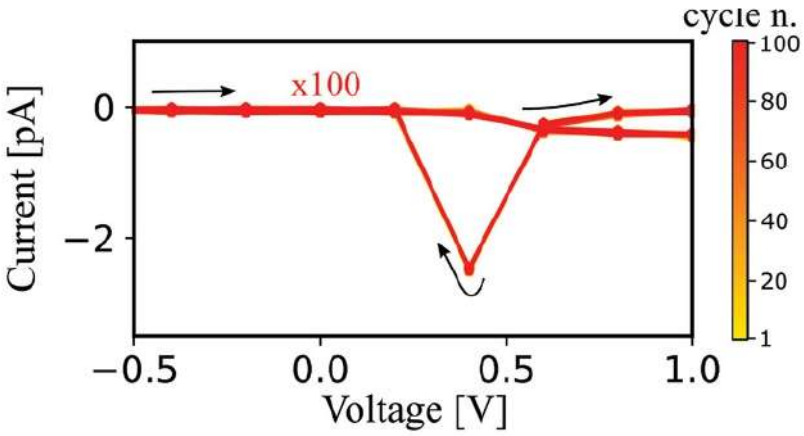
The IV curves obtained during CV tests, allowed us to detect the samples that revealed promising electrochemical behaviour. Figure 28 reports typical curves, where the arrows illustrate the voltage sweeping direction: in 28a is shown a device where no remarkable ionic currents can be identified, so that leakage and capacitive contributions, which cause the symmetrical narrow hysteresis around the y axis, are found to predominate. On the other hand, the plot shown in Figure

28b, which was typical for the samples of stack $CuTe_2Ge/Ta_2O_5/W$, features a clear current peak, that cannot be assigned to either leakage or capacitive currents.

REDUCTION PEAKS IN Ta_2O_5 CBRAM SAMPLES Indeed, we expect to observe oxidation and reduction peaks in those devices where enough ions are successfully injected as consequence of redox processes. The appearance of a current peak over the falling ramp of the voltage biasing a DUT can be associated to a reduction reaction, taking place inside the solid electrolyte.



(a) A typical voltammogram, featuring a clear current peak during the negative scan rate.



(b) At lower resolution, a current peak over 100 cycles, where the curves perfectly overimpose.

Figure 29 – Cyclic Voltammetry curves obtained for two different samples, showing a clear current peak when sweeping towards negative voltages (arrows show the sweeping direction).

A high resolution plot, for a 5nm-thick sample, is shown in Figure 29a. The current peak appears when sweeping through V_b's falling ramp, at approximately 200mV. Figure 29b shows 100 CV cycles, where a lower resolution was adopted during tests. It can be seen that the curves overlap each other, confirming the repeatability of the process (in Appendix Section 4.7.1, supplementary curves in higher resolution).

Given the absence of a reference electrode, it is nontrivial to assign a specific reaction to the current peak. Aqueous solution potentials suggest that the reduction of Cu^+ ions to Cu is the most energetically favorable [110]. Although our (solid) electrolyte is far from an aqueous solution, some research groups hypostatized that water molecule might be incorporated inside the nano-pores of the Ta_2O_5 layer due to ambient humidity [100, 101, 109], therefore aqueous potentials might still hold true. Figure 30 shows likely reactions taking place in the examined stack, during a positive scan rate. Assuming the presence of some water molecules inside the switching layer, the top electrode oxidation reaction (ionization of copper atoms into Cu^+) could be sustained either by the reduction of water molecules or copper ions [100, 101] at the Ta_2O_5/W interface. In particular, the reduction of water might be responsible for Cu^+ generation when no prior copper cation is present, and afterward concur with the reduction of copper ions at the bottom interface. Upon negative scan rate, the reactions illustrated in Figure 30 are expected to take place at the opposite interfaces: metallic copper oxidation at the Ta_2O_5/W, while copper reduction at the $CuTe_2Ge/Ta_2O_5$.

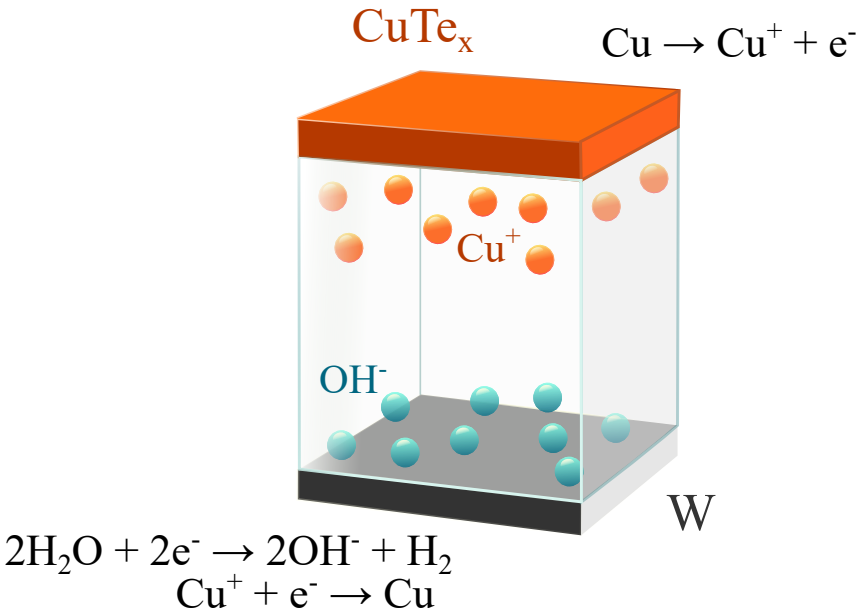


Figure 30 – Presumed redox reactions taking place during a positive scan rate: oxidation occurs at the top electrode while reduction at the bottom. Water molecules are possibly helping sustain the reduction reactions at the lower electrolyte/inert electrode interface.

Hence, our interpretation is that the current peak which appears in our CV tests results from the reduction of copper ions at the $CuTe_2$ top electrode. Although no oxidation peak appears when sweeping to positive voltages, this result agrees with what was reported in the literature for samples featuring high density Ta_2O_5 layers, where it was found that the oxide density trades with electrochemical currents. In these cases, no oxidation peak was observed when the electrolyte density was above $7.1 g cm^{-3}$ [105], a value that closely approaches the one of our samples.

2.4.3 *Impact of voltage, scan rate, area and temperature*

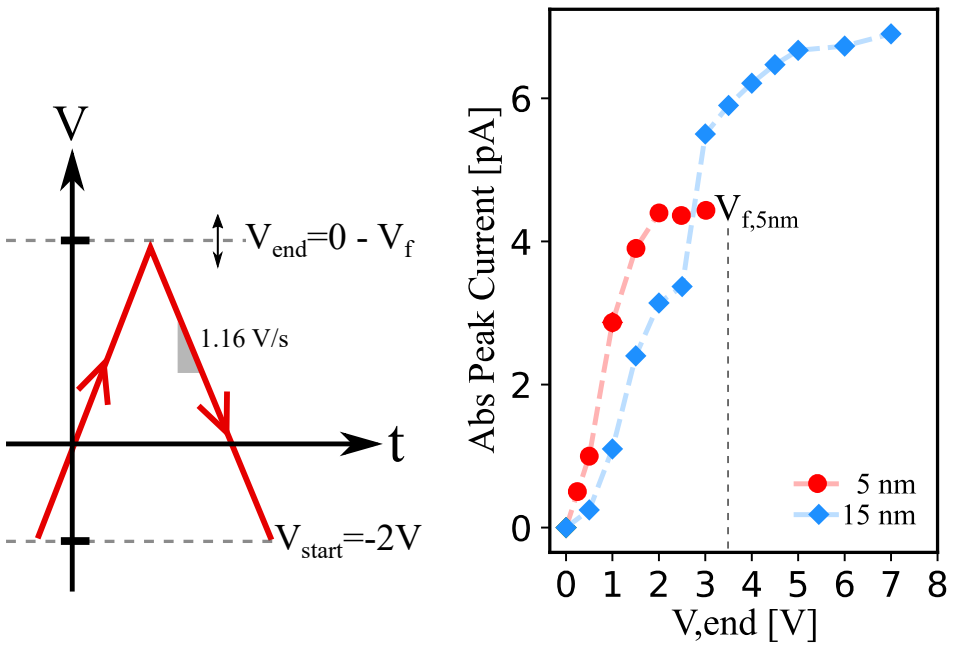
In order to confirm the electrochemical nature of the current peak observed in our samples, we proceeded to perform various further characterization; namely, we analyzed the impact of oxidation voltage amplitude, scan rate, sample area and temperature on the revealed current.

IMPACT OF OXIDATION VOLTAGE Figure 31 reports our analysis on the impact of the oxidation voltage on the current peak amplitude. Different tests are carried out by varying the oxidation potential from test to test, while keeping the reduction potential constant. If the current peak is indeed a result of copper ion reduction at the top electrode, it is expected to scale with the oxidation voltage.
The voltage biasing the cell is illustrated in Figure 31a, while Figure 31b shows the resulting curves, each representing a different oxide thickness: namely 5nm and 15nm. For the 5nm case, V_{end} is limited to 3.5V, as the device undergoes electroforming at 4V. A forming process was however not observed for the 15nm sample, up to $V_{end} = 8V$. For either case, the peak amplitude was found to increase with the maximum positive voltage, while no current peak was obtained for $V_{end} = 0$ (the measured current approaching the noise level). For $V_{end} > 0$, the relationship between voltage and current is initially close to linear (up until $V_{end,5nm} = 2V$ and $V_{end,15nm} = 4V$ respectively), with the 5nm case exhibiting a higher slope. By further increasing V_{end}, the curves reach a plateau.
Therefore, a clear relationship between the current peak and the oxidation voltage emerged from our tests. Our results confirm the electrochemical nature of the process observed, and indicate that the cation content inside the oxide can effectively be increased by heightening the oxidation voltage. The curve flattening, observed for higher potentials, might indicate that a maximum solubility of copper ions inside the Ta_2O_5 layer is reached. Although no oxidation peak is detected during positive scan rates, some oxidation processes must be taking place nonetheless, providing copper cations which are later reduced when the scan polarity is reversed.

IMPACT OF THE SCAN RATE We proceeded in our characterization by performing further CV tests at various scan rates, ranging from $160mVs^{-1}$ to $5Vs^{-1}$. Figure 32a shows an overlap of six curves, each obtained at a different sweep speed. It can be observed that, by increasing the scan rate, the current peak amplitude also increases, and the voltage at which the peak is obtained drifts to more negative values. The shift in peak position is typically observed for irreversible electrochemical processes, and thus we identified our reactions as such. The Randles-Sevick equation governing an irreversible process declines as [106]:

$$I_p = 0.4958(Fn)^{3/2}(RT)^{-1/2}Ac_0(\alpha D\nu)^{1/2} \tag{2}$$

where I_p is the peak current, F the Faraday constant, n the number of electrons exchanged in the reaction, R the gas constant, T the absolute temperature, A the cell's electrode area, c_0 the concentration coefficient, α the transfer coefficient, D the diffusion coefficient and ν the scan rate. Equation (2) dictates a linear trend



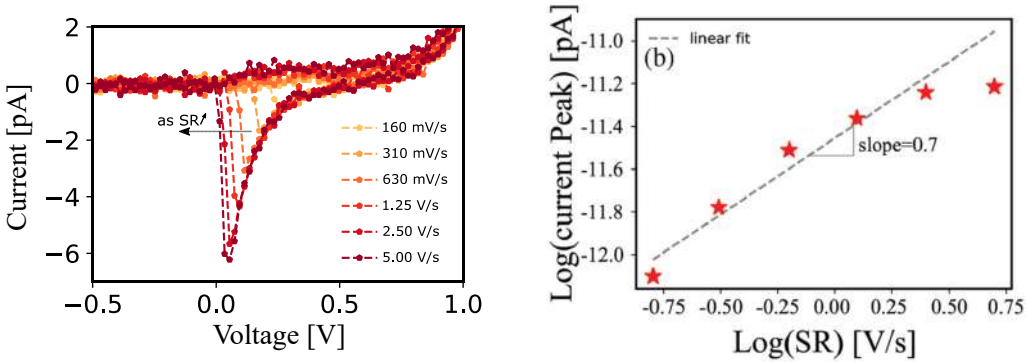
(a) Biasing voltage ranges: the oxidation voltage, V_{end}, was varied between 0 and the forming value V_f, while the reduction (starting) potential, V_{start}, was kept fixed at −2V.

(b) Current peak amplitude (absolute value) versus V_{end}. Different slopes can be observed for different thicknesses. A saturation trend bends the curves for both cases towards more positive values.

Figure 31 – Analysis of the impact of the oxidation voltage on the current peak.

between the peak current I_p and the square root of the scan rate ν. In order to verify that this relationship is satisfied in our samples, we plotted a log–log plot of current peak amplitudes versus the scan rate, shown in Figure 32b. A linear fit was computed, resulting in an extrapolated slope of 0.7, which can be compatible with currents resulting from redox reactions (Equation (2) expresses a slope of 0.5).
It is worth pointing out that, although some low capacitive currents might be present, our results do not favor this interpretation. In fact, in such a case the slope is equal to 1, and the current features a broad hysteresis (box-shaped curve), which is not the case for the curves reported in Figure 32a.
Finally, the trend of the current to the scan rate appears to deviate from the linear fit when the voltage sweep is fastened. The bend of the curve might be explained through relatively low reaction kinetics, which cannot keep up to the pace when the potential is rapidly changing.

IMPACT OF AREA The Randles-Sevcik Equation expressed by (2) also dictates a linear relationship between the device area and the current peak. Therefore, we measured the trend for our samples by comparing CV results, performed at the same voltage window and scan rate, for devices with different areas.
The sample area was first varied by selecting a different diameter size, in the range of 0.3µm to 1.7µm, as shown according to *Method 1* of Figure 33. It can be seen that we did not appreciate a (peak) current variation when varying the sample's diameter. Hence, we proceeded with *Method 2*, where the area was enlarged by



(a) In different colors, CV curves at various sweep rates.

(b) Current peak amplitude versus the scan rate, with linear fit.

Figure 32 – Analysis of the impact of the scan rate on the current peak.

physically connecting multiple devices in parallel. As shown, in this configuration we were able to measure a linear area-peak current trend, accordingly to our expectations.

Our interpretation is shown in Figure 34: the three scenarios show, respectively, a single cell, a single cell with a larger diameter, and n cells connected in parallel. $Vdd - Vss$ is the voltage drop across the sample(s), which triggers redox reactions so that charged species are released inside the cell's switching layer. The unitary charge delivered by a single cell is indicated as Q_u, and the overall charge stored by the system as Q_s. Our hypothesis is that the reactions taking place in a device are limited to a restricted portion of the electrolyte, in a similar manner to the resistive switching effect, so that increasing the diameter alone does not play a role in increasing the overall stored charge. On the other side, when connecting n cells in parallel, it is possible to increase the amount by a factor n. Hence, we conclude that the energy storage inside our samples shows *local* trait.

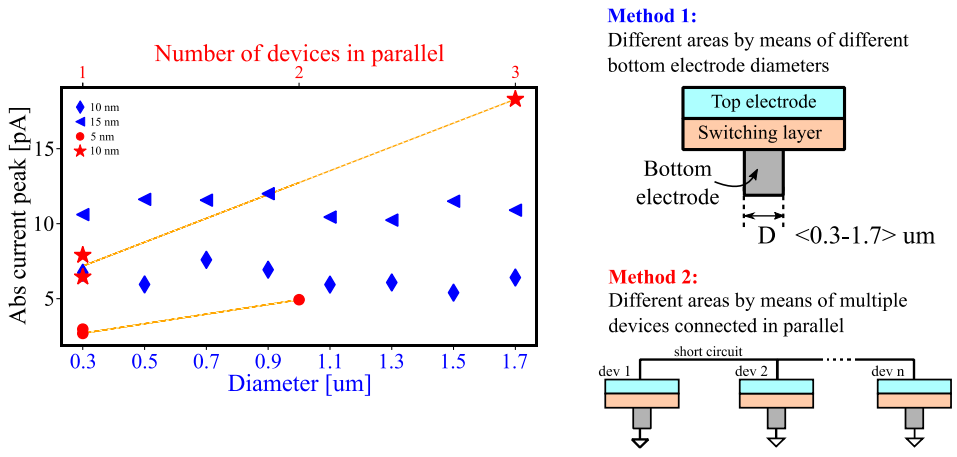


Figure 33 – Trend of peak current versus area, increased either by means of a larger bottom electrode diameter (Method 1, in blue) or by multiple devices connected in parallel (Method 2, in red).

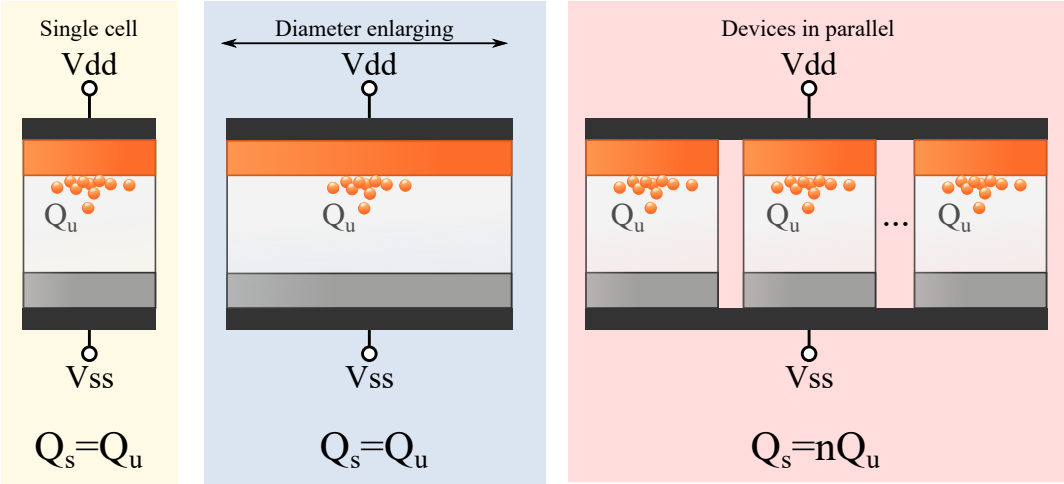


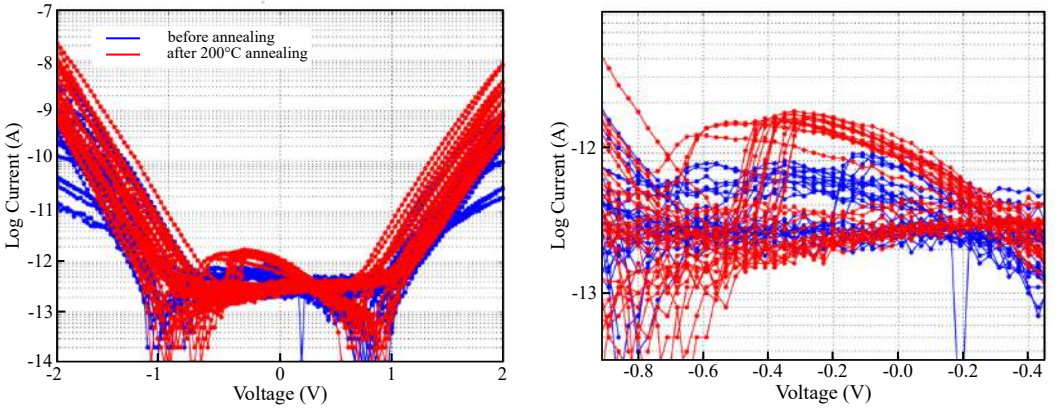
Figure 34 – Interpretation of the current - area dependence observed in our samples. When the area of a single cell is increased by enlarging the cell diameter, the total charge remains unchanged, as consequence of a limited portion of the electrolyte being electrochemically active. On the other hand, when multiple samples are connected in parallel, the number total stored charge can be increased.

IMPACT OF TEMPERATURE In an attempt to stimulate greater ion injection though thermal effect, we annealed a wafer at 200°C for 30 minutes. Upon completion of the process, we waited for the wafer to cool down and then proceeded to perform CV tests. The sampled tested were 5nm thick, and of different diameter size (0.3µm to 1.7µm), in order to see if an area dependence would be obtained after the annealing process. Figure 35a shows the resulting curves, superimposed, each belonging to a device of different diameter; in blue, the curves before annealing while in red, afterwards (the tests were performed as soon as the wafer reached room temperature). Figure 35b shows a zoomed view of the current peak, where the curves in red display a clear amplitude increase (~ double) compared to the ones in blue. However, no difference in amplitude can be appreciated between samples of different diameter. Therefore, the annealing process looks effective in boosting the quantity of injected charge, although the mechanism still appears to be involving only a limited portion of the oxide.

SUMMARY TABLE Table 3 shows a summary of the CV tests presented in this Section. The different parameters considered in our analysis are reported, along with the observed effect on our samples.

2.5 DIFFUSION AND CONCENTRATION COEFFICIENT EXTRAPOLATION

The experimental results presented in previous Section 2.4.3 demonstrated that the currents peaks observed during CV tests are indeed product of faradaic currents. Therefore, as a next step, we deepened our investigation by deriving the diffusion and concentration coefficients characterizing the reactions taking place in our samples.



(a) In blue, curves obtained before annealing while in red, after treatment. Same color curves are associated to samples of different diameter, in the range 0.3µm to 1.7µm.

(b) Zoomed view of the current peaks, before (blue) and after (red) annealing. An amplitude increase of a factor 2 (on average) can be appreciated after treatment. No area variation can still be observed.

Figure 35 – Analysis of the impact of the temperature on the current peak.

The recombining charge during the reduction process, Q_{Cu^+}, can be derived by discrete time-integration of the current during the negative voltage scan:

$$Q_{Cu^+} = dt \left| \sum_{0<n\leqslant N} i_n \right| \tag{3}$$

where δt is the time interval between each sample, i_n the n^{th} current sample value and N the total number of samples. In order to exclude tunneling current contributions, the voltage interval of interest was restricted to $[-0.2, 0.6]V$, where no tunnelling effect is observed.
The injected moles, N_{Cu^+}, can be derived with Faraday's law:

$$N_{Cu^+} = \frac{Q_{Cu^+}}{nF} \tag{4}$$

where n is the oxidation number, assumed equal to 1 (for Cu^+ ions), and F the Faraday constant. At this point, the concentration coefficient, c_{Cu^+} can be evaluated. Considering our sample's cylindrical geometry:

$$c_{Cu^+} = \frac{N_{Cu^+}}{V_{Ta_2O_5}} = \frac{N_{Cu^+}}{t\pi(d/2)^2} \tag{5}$$

where $V_{Ta_2O_5}$ is the volume of the (oxide) electrolyte, t its thickness and and d the diameter of the bottom electrode.
Figure 36a reports the concentration coefficient versus the scan rate: it can be seen that c_{Cu^+} reasonably drops when increasing ν, in accordance with the Randles-Sevcic equation ((2)). This trend can be explained with the fact that less and less time is available for reactions to generate products when the potential is swept at an increasingly faster pace.

Parameter	Method	Effect
Oxidation voltage, V_{ox} (p. 51)	V_{ox} varied between 0 and forming voltage V_f.	I_{pk} found to scale with V_{ox}. Thickness affects the dependency. Saturation trend sets in for the upper range of V_{ox}.
Scan rate, ν (p. 51)	ν was ranged between $[0.16-5]Vs^{-1}$ (5 values tested).	I_{pk} found to scale with ν; a slope of 0.7 was extrapolated in log-log plot.
Area, A (p. 52)	Device with different diameter tested, in the range $[0.3-1.7]\mu m$ (8 different diameters tested). A was also enlarged by connecting 2 and 3 samples in parallel.	No change in I_{pk} observed upon varying the device diameter alone. However, I_{pk} was found to scale linearly with A when multiple devices tested in parallel.
Temperature, T (p. 54)	Devices were tested at room temperature, and after annealing at 200^oC.	An average increase of a factor of 2 in I_{pk} observed after annealing, when comparing to the peak amplitude beforehand.

Table 3 – Impact of the various parameters considered in our analysis on the current peak amplitude, I_{pk}.

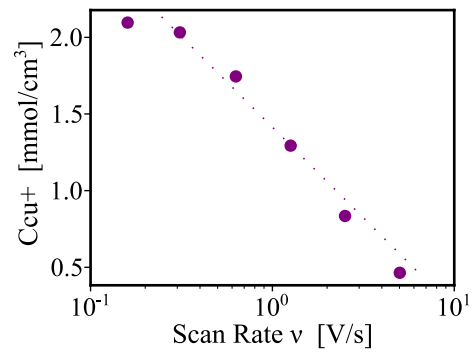
Once the concentration coefficient is known, the diffusion coefficient can be extrapolated using Equation (2):

$$D = \frac{RT}{\alpha\nu(Fn)^3}\left(\frac{I_p}{0.4958Ac_{Cu^+}}\right)^2 \tag{6}$$

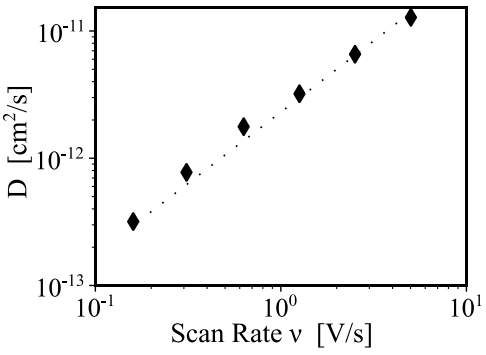
where α is the charge transfer coefficient, and has been taken equal to 0.5 [105]. The diffusion coefficient versus the scan rate is plotted in Figure 36b, while Figure 36c shows D versus the concentration coefficient, c_{Cu^+}. It can be seen that the trend of the diffusion coefficient versus the scan rate is opposite to that of the concentration coefficients: this result is reasonable, as diffusivity increases when the density of (Cu^+) ions decreases.
Our extrapolations are found in agreement with reports from the literature for similar Ta_2O_5 based RRAM stacks [105]. Finally, the energy storage capability was evaluated through:

$$E_{SR,i} = dt\left|\sum_{0<n\leqslant N} i_n \cdot V_n\right| = dt \cdot P \tag{7}$$

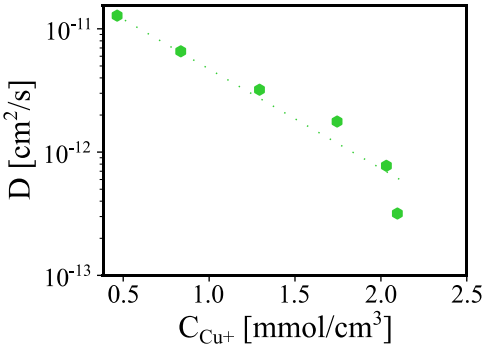
where V_n are the values of the voltage corresponding to the n^{-th} current sample, i_n, and P the released power. Figure 36d reports the evaluated energy and power densities versus the scan rate. It can be seen that the energy density is lower when the scan rate is higher, in which condition less ions are present, and thus less charge can be stored inside the electrolyte.



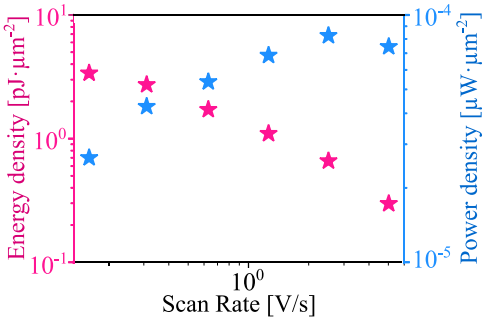
(a) Concentration coefficient, c_{Cu^+}, versus the scan rate, ν.



(b) Diffusion coefficient, D, versus the scan rate, ν.



(c) Diffusion (D) versus concentration coefficient (c_{Cu^+}).



(d) Energy and power density per device, versus the scan rate.

Figure 36 – Extrapolated electrochemical parameters.

2.6 CONSIDERATIONS AND PERSPECTIVES

The analysis conducted on our RRAM samples, for novel *in-memory energy* use, allowed us to shine some light on their fundamental characteristics, which can be summarized by the following considerations:

1. The proposed technology allows to operate a device simultaneously as memory and energy storage element, whenever the cell is storing a logic 0 (RRAM in IRS or HRS). In fact, electrochemical reactions are expected to take place either before the creation, or after the dissolution of the conducting filament. Figure 37a illustrates the operating regions, in terms of qualitative voltage and current domains, of a such dual-behavior RRAM based device.

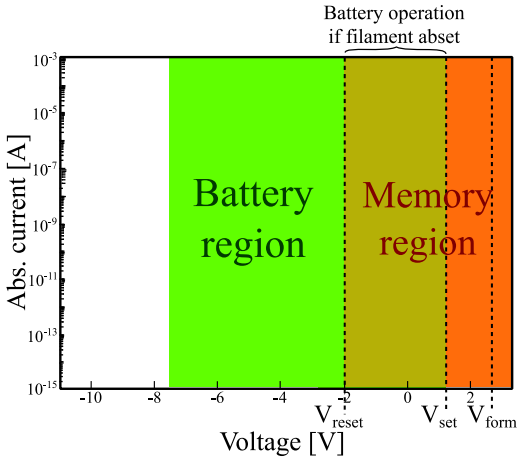
2. Our study put into evidence that the faradaic processes happening inside our samples seem not to involve the whole volume, as a linear trend of the

current peak with the area could not be satisfied when enlarging the device diameter, but was instead verified when connecting multiple devices in parallel. Consequently, the energy storage mechanism appears to be a "local" phenomenon, possibly limited to the switchable oxide region of the solid electrolyte. As a result, energy capability should be increased by maximizing the number of devices connected in parallel, meaning that technological downscaling would be beneficial to boost both memory and energy density.

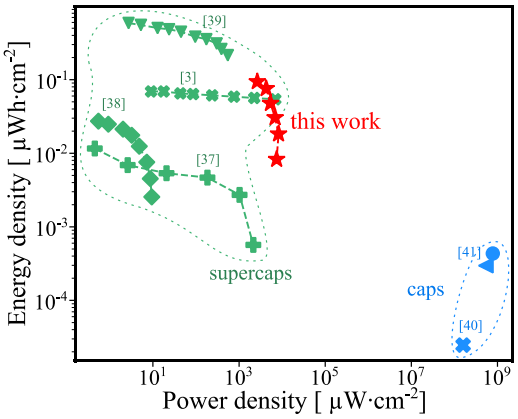
3. The estimated energy density achievable peaks, for a standalone cell, around $3.5pJ\mu m^{-2}$, at the slowest scan rate $\nu = 160mVs^{-1}$. The maximum power density amounts to $80pW\mu m^{-2}$, obtained at $\nu = 5Vs^{-1}$. Figure 37b shows a Ragone plot, which benchmarks and compares our devices to other SOA integrated energy storage elements. Remarkably, the energy and power densities extrapolated for our samples appear to rival with those of integrated planar supercapacitors. Figure 37c reports stored energy versus sample area, highlighting that the energy could theoretically be linearly increased with the number of devices connected in parallel, up to values comparable to those of integrated planar supercapacitors. Furthermore, the proposed technology (benchmarked at a minimum area of $0.07\mu m^{-2}$) provides the added advantage of being much more scalable, opening the possibility of realizing "deep granular supercapacitor" by means of RRAM elements.

2.7 CONCLUSIONS AND REMARKS

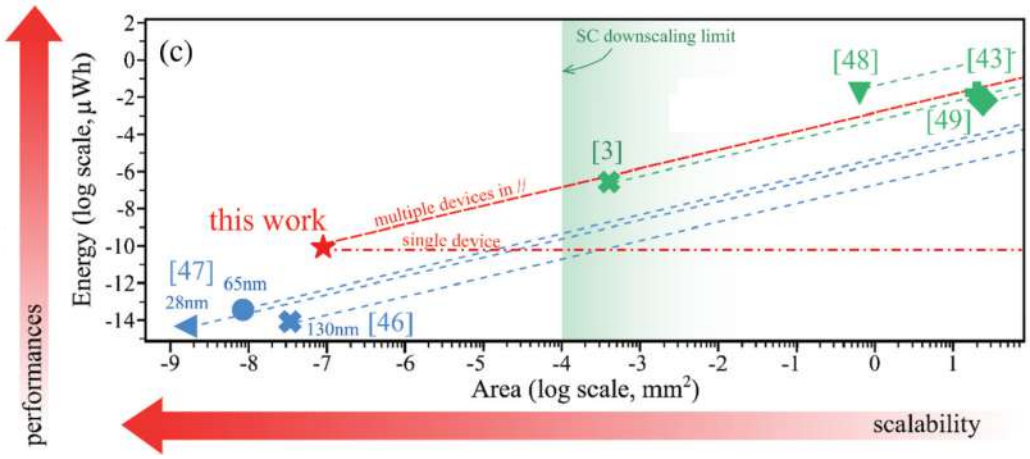
In this chapter, we explored the new concept of RRAM-based in-memory energy storage. We conducted preliminary electrochemical characterization on SOA CBRAM, by means of Cylic Voltammetry tests. Our study allowed us to confirm the faradaic nature of the processes taking place inside our samples, as wells as delineate some main properties, such as the local energy storage trait, where only a portion of a cell's electrolyte seems to be involved in ion exchanges. Consequently, the proposed technology offers the advantage of raising both the energy and memory density with downscaling. Future envisage-able designs would feature minimally sized cells, connected in parallel when operated as energy source, and placed in close proximity to the load.
The extrapolated energy and power densities compare with planar integrated supercapacitors, with the added benefit of being much more scalable, and highly compatible with CMOS fabrication.
In conclusion, our research expresses potential and could be a disruptive solution in the field of integrated energy sources. Moreover, it could be adopted in existing fields, for example IMC, adding the energy storage feature to broaden its advantages.
Nevertheless, our study is still at a preliminary stage, and further evaluation is demanded to deepen the understanding, and fully assess the traits of this new technology. In particular, future work should aim to quantify the output voltage, coulumbic efficiency and experimental energy and power density. As a next step, new design solutions where the memory and energy performances are optimized



(a) Qualitative voltage and current operating regions of a dual-behaviour RRAM device. An overlap region exists when the element is in HRS, acting simultaneously as a memory and battery cell.



(b) Ragone plot, comparing our work (in red) with planar supercapacitors (in green) and on-chip dielectric capacitors at various CMOS nodes (in blue).



(c) Area versus energy storage. Higher energy can be obtained by connecting multiple devices in parallel (extrapolated), up until values comparable with SOA supercapacitors. On the other side, the minimal size achievable by a battery-memory cell is comparable with integrated dielectric capacitors, making the technology highly scalable.

Figure 37 – Extracted specifications of our RRAM-based in-memory energy devices.

could be explored, in order to bring the concept of in-memory energy storage to the application level.

2.8 CHAPITRE 2 - RÉSUMÉ EN FRANÇAIS

Ce chapitre a présenté notre étude sur la caractérisation de la technologie RRAM comme un nouveau type de source d'énergie. Etant donné que son comportement mémoire est basé sur des réactions faradiques, nous avons réalisé une étude préliminaire sur la faisabilité d'une nanobatterie à base de RRAM.
Le principal outil utilisé pour caractériser les échantillons de mémoire était la voltamétrie cyclique, qui est largement adoptée en électrochimie afin d'inspecter les processus redox qui se déroulent à l'intérieur de l'électrolyte (solide). Nos tests ont révélé qu'un pic de courant peut être détecté lors de la rampe de tension négative, que nous avons associé à une réaction de réduction dans l'oxyde. Nous avons proposé des réactions possibles, en fonction de la composition matérielle des cellules analysées, ainsi que des rapports bibliographiques.

Afin de confirmer la nature électrochimique du pic de courant observé dans notre échantillons, nous avons procédé à diverses caractérisations complémentaires ; à savoir, nous avons analysé l'impact de l'amplitude de la tension d'oxydation, de la vitesse du scan de tension, de la taille d'échantillon et de la température sur le courant révélé. Nous avons observé que l'impact de ces variables répond en conséquence à ce que dicte l'équation de Randles-Sevcik pour un processus électrochimique irréversible. Donc, notre analyse nous a finalement permis de confirmer la nature électrique des processus en cours.

Notre étude nous a permis de dégager un certain nombre de traits fondamentaux de cette technologie, qui sont résumés ci-après.

1. La technologie proposée permet de faire fonctionner un dispositif simultanément comme élément de mémoire et de stockage d'énergie, chaque fois que la cellule stocke un 0 logique (RRAM en IRS ou HRS). En effet, des réactions électrochimiques sont censées avoir lieu soit avant la création, soit après la dissolution du filament conducteur. La figure 37a illustre les régions de fonctionnement, en termes de domaines qualitatifs de tension et de courant, d'un tel dispositif basé sur RRAM à double comportement.

2. Notre étude a mis en évidence que les processus faradiques se produisant à l'intérieur de nos échantillons ne semblent pas impliquer tout le volume, car une tendance linéaire du pic de courant avec la zone ne pouvait pas être satisfaite lors de l'élargissement du diamètre de l'appareil, mais a plutôt été vérifiée lors de la connexion de plusieurs appareils. en parallèle. Par conséquent, le mécanisme de stockage d'énergie apparaît comme un phénomène « local », éventuellement limité à la région d'oxyde commutable de l'électrolyte solide. En conséquence, la capacité énergétique devrait être augmentée en maximisant le nombre d'appareils connectés en parallèle, ce qui signifie qu'une réduction d'échelle technologique serait bénéfique pour augmenter à la fois la mémoire et la densité d'énergie.

3. La densité d'énergie estimée atteint des pics, pour une cellule autonome, d'environ 3,5 $pJ\mu m^{-2}$, à la vitesse de balayage la plus lente $\nu = 160\ mVs^{-1}$.

La densité de puissance maximale est de $80pW\mu m^{-2}$, obtenue à $\nu = 5Vs^{-1}$. La figure 37b montre un tracé Ragone, qui compare et compare nos appareils à d'autres éléments de stockage d'énergie intégrés SOA. Remarquablement, les densités d'énergie et de puissance extrapolées pour nos échantillons semblent rivaliser avec celles des supercondensateurs planaires intégrés. La figure 37c rapporte l'énergie stockée par rapport à la surface de l'échantillon, soulignant que l'énergie pourrait théoriquement être augmentée de manière linéaire avec le nombre d'appareils connectés en parallèle, jusqu'à des valeurs comparables à celles des supercondensateurs planaires intégrés. De plus, la technologie proposée (évaluée à une surface minimale de 0,07 μm^{-2}) offre l'avantage supplémentaire d'être beaucoup plus évolutive, ouvrant la possibilité de réaliser un "supercondensateur granulaire profond" au moyen d'éléments RRAM.

En conclusion, notre recherche exprime un grand potentiel et pourrait être une solution de rupture dans le domaine des sources d'énergie intégrées. De plus, il pourrait être adopté dans des domaines existants, par exemple l'In Memory Computing (IMC), en ajoutant la fonction de stockage d'énergie pour élargir ses avantages.
Néanmoins, notre étude en est encore à un stade préliminaire et une évaluation plus approfondie est requise pour approfondir la compréhension et évaluer pleinement les caractéristiques de cette nouvelle technologie. En particulier, les travaux futurs devraient viser à quantifier la tension de sortie, l'efficacité coulumbique et les densités expérimentales d'énergie et de puissance.
Dans une prochaine étape, de nouvelles solutions de conception où les performances de la mémoire et de l'énergie sont optimisées pourraient être explorées, afin d'amener ce nouveau concept au niveau de l'application.

Part III

INTEGRATED CIRCUIT DESIGN

3

CAPACITOR-BASED WRITING PROTOCOL

This chapter presents a novel concept for the programming a RRAM cell, based on charge capacitor, as opposed to a constant voltage source. The key idea lies in controlling the energy that is delivered to the memory element, which is bounded by the charge stored inside the capacitor. The proposed method is then compared to the state of the art, highlighting benefits and limitations. An alternative, dual approach, where the programming element is a charged inductor, is also presented in Section 3.7.

3.1 CHARGE-BASED SWITCHING CONCEPT

When RRAM are programmed following the standard protocol, i.e. by voltage-biasing the cell over a fixed time, a considerable amount of excess energy is, on average, consumed. This waste, or *over-programming*, is caused by the fact that the cell continues to unnecessarily sink current even after reaching its low resistive state. As a consequence, the writing process is afflicted by a low energy efficiency [60], as long as being damaging to key performances like reliability and endurance [111].

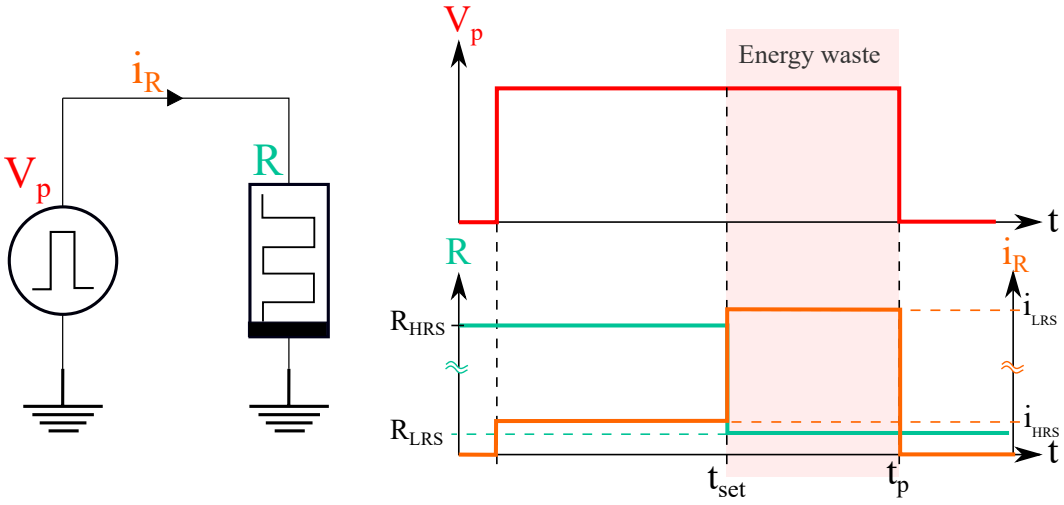
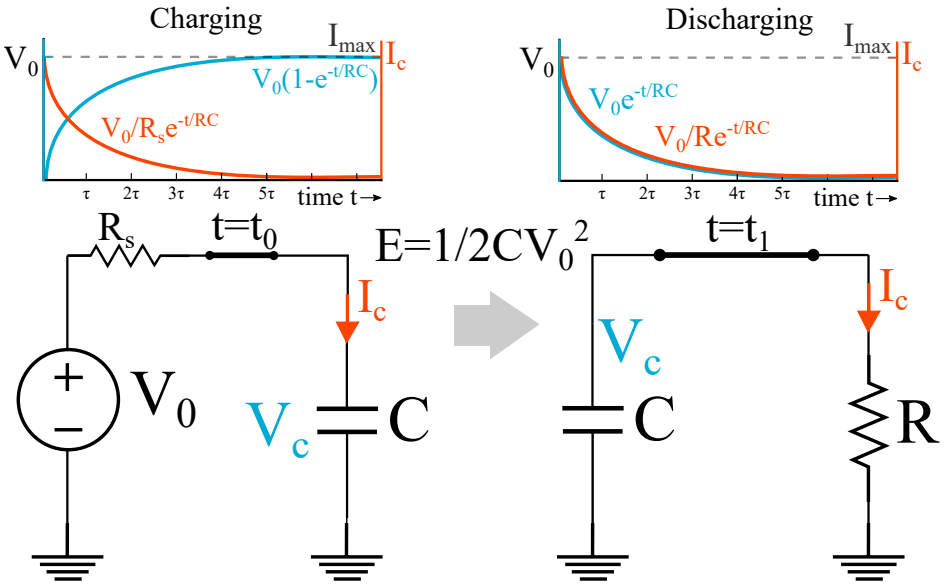


Figure 38 – RRAM set operation by means of a voltage pulse, where energy waste occurs whenever the transition to the LRS occurs before the falling edge of the pulse.

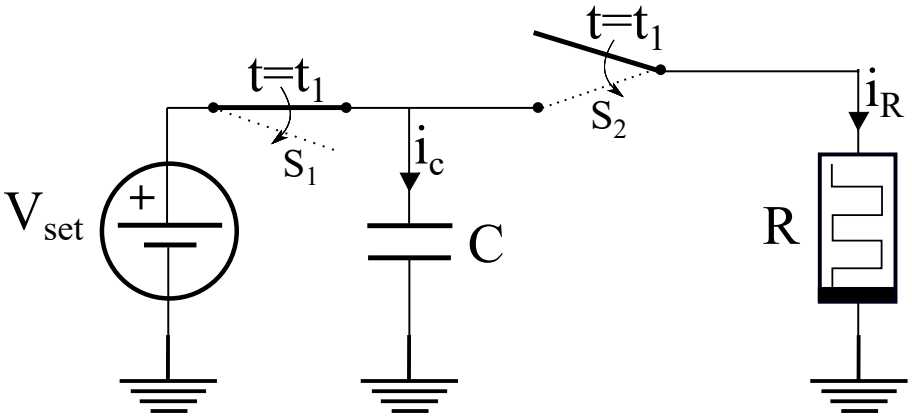
Figure 38 illustrates the problematic. The programming source, V_p, is a pulsed voltage of amplitude A and time duration t_p. It positively biases a RRAM load R, initially in HRS, in order to induce a set event. The set energy, i.e. the energy amount necessary to trigger a high to low resistance transition, suffers from a high variability, both from cell to cell and cycle to cycle [60]. Consequently, the programming source must abide the requirements of the most energy-demanding cells, in order to guarantee a high switching probability. In other words, the time duration

of the programming pulse, t_p, should be equal to the longest cell switching time, $t_{set,MAX}$. It thus follow that all the elements that switch earlier, at $t_{set} < t_p$, will be highly conductive and drain considerable extra energy over a time interval $t_p - t_{set}$.

Our approach to ease the problematic relies on the use of a charged capacitor to perform the programming operation, instead of a Constant Voltage Source (CVS). This way, the programming energy can be *dosed* through the amount of charge stored inside the capacitor, enabling higher efficiency and lower device-damage risk, regardless of the switching time dispersion. Figure 39a depicts the proposed procedural sequence, where a capacitor is initially charged by a CVS, and later switched onto a generic resistive load R.



(a) Capacitor-based programming sequence. The programming capacitor, C, is initially charged to programming voltage V_c and later discharged onto resistor R in order to perform the resistive switching operation.



(b) Proposed Constant Charge Source (CQS) approach, where switches S_1 and S_2 control the programming flow so that C is discharged onto the memory cell, R, once fully charged to voltage V_{set}.

Figure 39 – Capacitor-based memory programming (CQS).

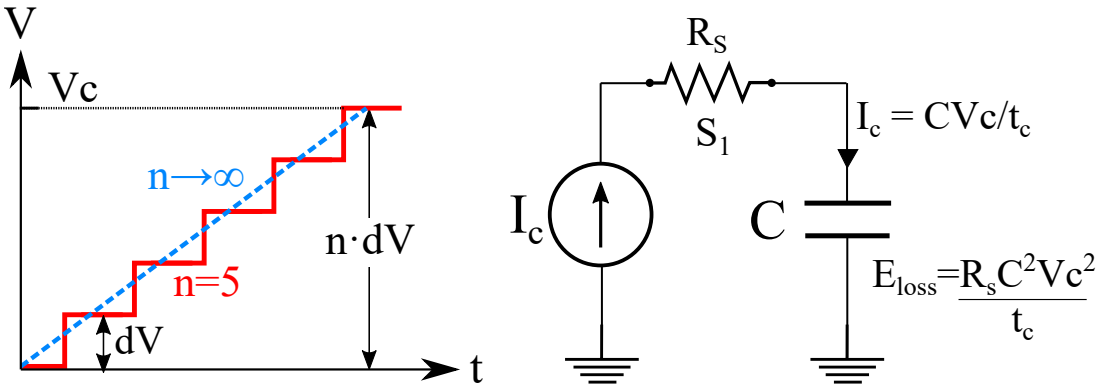
The energy supplied by programming capacitor C can be expressed as:

$$E = \frac{1}{2}C(V_c - V(\infty))^2 = \frac{Q^2}{2C} \tag{8}$$

where C is the capacitance of the programming capacitor, V_c the charging voltage, $V(\infty)$ the residual voltage at the end of the discharge phase ($V(\infty) = 0$ if the capacitor is fully discharged onto the load), and Q the stored charge. The amount of charge, Q, that is stored inside the programming capacitor will set and upper-bound to the energy delivered to the load.
Figure 39b illustrates a simple circuit which could be used to accomplish the set operation for a RRAM. Switch S_1 is closed until $t = t_1$, allowing the charging of programming capacitor C to voltage V_{set}, while switch S_2 is open. At $t = t_1$, the switch positions are reversed, and S_2 couples C onto the RRAM cell, allowing the high-to-low resistance transistion to take place. The proposed method will henceforward be referenced as CQS, as opposed to the CVS standard.

3.2 CHARGING EFFICIENCY

The circuit illustrated in Figure 39b, where the programming capacitor is charged by a voltage source, is arguably flawed. In fact, this configuration is intrinsically inefficient, as half of the energy delivered by the charging source will be dissipated onto switch S_1's resistance. The heat losses might be minimized by stepping the charging voltage according to the principle of adiabatic charging [112–114]. Figure 40a illustrates this concept: the charging efficiency is found to linearly increase with the number of steps, so that the (ideally) lossless case can be obtained when $n \to \infty$. This translate into charging the capacitor through a voltage ramp, which can be practically achieved using a current source, as shown in Figure 40b. In this situation, a near-100%-efficient charging process can be obtained; the residual losses being caused by the heat dissipated over switch S_1 resistance R_s, over charging time t_c.



(a) Charging to V_c through n intermediate steps allows to reduce the charging losses by a factor of n.

(b) Ideal capacitor charging. The residual losses are caused by the heat dissipation through the switch parasitic resistor R_s.

Figure 40 – Adiabatic charging: gradual source voltage increase allows to improve efficiency.

3.3 COMPARISON BETWEEN CVS AND CQS

In order to validate our approach, and to benchmark its benefits in comparison to the SOA, we designed a suitable set behavioural model for reliable simulation in CQS condition. We tuned our model with experimental data, and successively compared our extrapolated results between CQS and CVS.

3.3.1 *Equivalent RRAM circuit for a set process*

The concept of programming a RRAM through a charged capacitor has been first reported by Zhang et al [115]. Relying on the literature [101, 102], they built a set behavioral model that is both voltage and charge sensitive. Here, we introduce a new RRAM set equivalent circuit, which is voltage and energy sensitive. Our proposal brings the advantage of easing the extrapolation, from experimental data, of the switching parameters:

- V_{th}: activation voltage for the redox reactions responsible for filament formation inside a RRAM cell [101, 102]. It is equal to the minimum set voltage for a given technology.
- E_{th}: threshold energy, i.e.the minimum energy dissipation that must occur over a memory cell before the set operation is accomplished.

Consequently, a set operation can take place once the voltage over a RRAM cell, V, and the energy dissipated onto it, E, satisfy the two conditions:

- $V > V_{th}$
- $E > E_{th}$

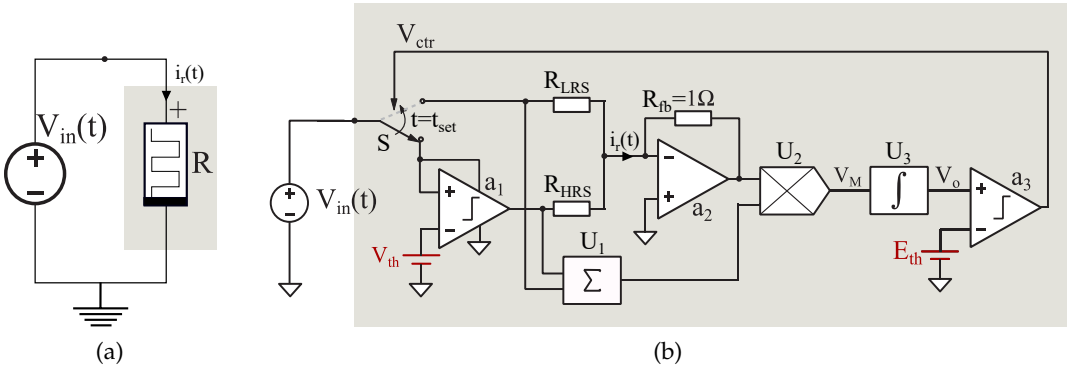


Figure 41 – Equivalent voltage and energy sensitive circuit for a set process.

Figure 41 show an electrical equivalent circuit which accommodates the aforementioned conditions. Figure 41b expands the RRAM component in Figure 41a. Initially, the RRAM is in HRS, so that the input voltage $V_{in}(t)$ drops over the large off-state resistance R_{HRS}. $V_{in}(t)$ is buffered by comparator a_1 to its output if $V_{in}(t) > V_{th}$, as its non-inverting terminal is biased to V_{th}. Successively, the current sinked by the cell, $i_R(t)$, is converted into a voltage of equal amplitude by transimpedance amplifier a_2 (of unitary gain). Summer U_1 outputs the voltage

drop over the sample, $V_{in}(t)$, regardless its logic state. Analog multiplier M performs multiplication of $V_{in}(t)$ and $i_R(t)$, thus delivering the instant power at its output:

$$V_M = V_{in}(t) i_R(t) dt = P(t) \tag{9}$$

Finally, the energy dissipated is then obtained at the output of ideal inverting integrator U_3, which time-integrates the instant power. As a result, V_o:

$$V_o = \int_0^T P(t) dt = \int_0^T V_{in}(t) i_R(t) dt = E(T) \tag{10}$$

where T is the integration time, and $E(T)$ the energy consumed by the memory cell after a time T. When V_o hits E_{th}, comparator a_3 output rises from gnd to Vdd, and the input switch S closes onto R_{LRS}, accomplishing the set operation. The presented model holds a strong assumption, by replacing the RRAM load as a constant resistor, and thus assuming ohmic conduction is predominant in both HRS and LRS. While this holds true after the set operation has taken place, it is usually not the case prior to filament formation, where non-linear tunneling effects are found to predominate. Nevertheless, the approximation is justifiable in our framework, where the aim is to provide a first-order estimation of the energy consumption for comparison with different programming techniques.

3.3.2 *Model calibration with experimental data*

The equivalent circuit of Figure 41b well-adapts to SPICE simulation, previous calibration of parameters R_{HRS}, R_{LRS}, V_{th} and E_{th}, which are technology related. Indeed, these values can be easily extracted from standard electrical characterization in CVS. V_{th} can be put equal to the minimum set voltage that allows a satisfying success rate at a given pulse length. The threshold energy E_{th} can be derived from:

$$E_{th} = t_s(V_{set}) \frac{V_{set}^2}{Z_{HRS}(V)} \tag{11}$$

where V_{set} is the amplitude of the pulse, $t_s(V_{set})$ the cell voltage-dependent switching time and $Z_{HRS}(V)$ the voltage-dependent impedance of the sample in HRS. Equation (11) expresses the energy consumption over the sample until the filament forms and the cell transits to its LRS. Although statistical dispersion is expected for HRS and LRS resistance values, we have set:

$$Z_{HRS} = \overline{R_{HRS}} \tag{12}$$

where $\overline{R_{HRS}}$ is the mean value of the HRS resistance dispersion. This approximation, made for the sake of simplicity, is intended as a first order estimation of the actual consumed energy in HRS, and can be justified by the fact that the spread of the switching times (and thus energies) already reflects cell-to-cell variability of

the R_{HRS} values.
The total energy dissipated during the set operation is calculated as the sum of the necessary energy to set, E_{th}, and the portion that is wasted after the cell has reached its LRS but is still being biased by the programming pulse, E_w:

$$E_{tot} = E_{th}(t_s) + E_w(t_s, t_p) \tag{13}$$

where $E_{w(t_s,t_p)}$ is the switching time and pulse duration dependant energy waste, which can be expressed as:

$$E_w = \frac{V_{set}^2(t_p - t_s)}{\overline{R_{LRS}}} \tag{14}$$

where $\overline{R_{LRS}}$ is the mean LRS. Finally, by substituting expressions (11) and (14) in equation (13), the overall energy consumption declines as:

$$E_{tot} = \frac{V_{set}^2 t_s}{R_{HRS}} + \frac{V_{set}^2(t_p - t_s)}{R_{LRS}} \tag{15}$$

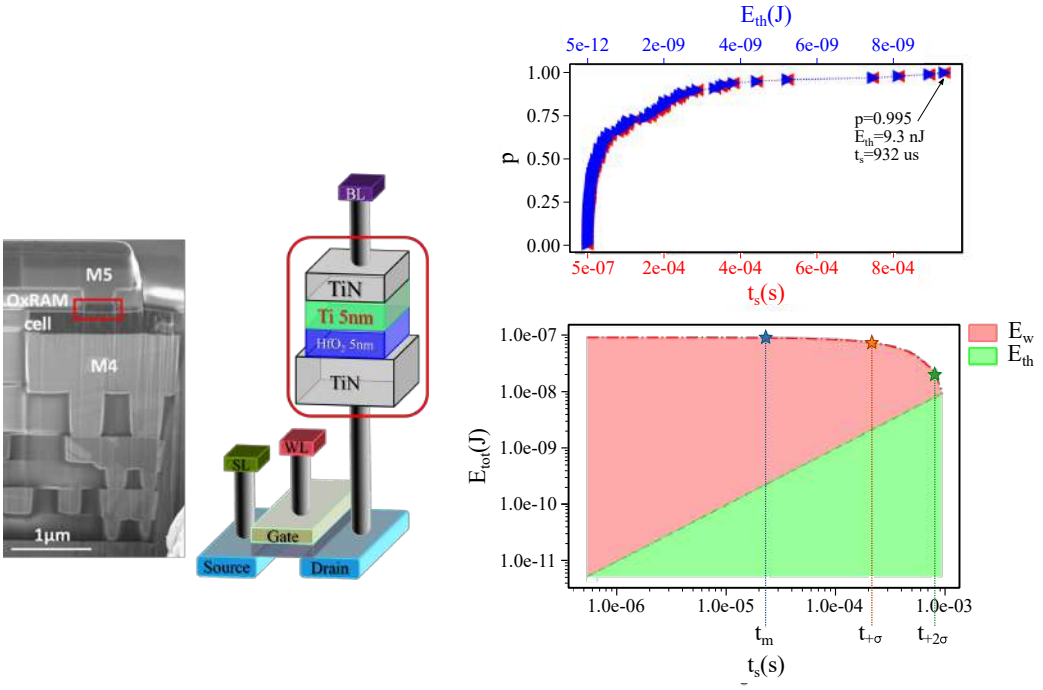
Whereas the programming efficiency can be quantified as:

$$\eta = \frac{E_{th}}{E_{th} + E_w} \tag{16}$$

CHARACTERIZATION IN CVS We performed electrical characterization on 1kb $TiN/HfO_2/Ti-TiN$ samples, in 1T1R cell configuration, where the memory array was fabricated as BEoL, on top of the access transistor, integrated in bulk 130nm CMOS process. The test setup included a pulse generator to induce the set operation, thus resembling a CVS source, of amplitude fixed at $V_{set} = 1V$. Switching times were collected by marking the instant the current over a sample hit the compliance value. Figure 42a shows a TEM picture of the tested stack.
The top of Figure 42b shows, in red, the switching times and, in blue, the switching energies versus the cumulative probability of set. E_{th} values were derived from the measured $t_s(1V)$ according to equation (11), where $\overline{R_{HRS}} = 100k\Omega$. The switching times were found to range from 500ns to 932µs, where the latter value corresponds to the highest switching probability, equal to a Success Rate (SR) of 99.5%. The bottom of Figure 42b shows the overall energy consumed for the set operation, E_{tot}, versus the switching time t_s, quantified using expression 15, where $V_{set} = 1V$, $\overline{R_{HRS}} = 100k\Omega$, $\overline{R_{LRS}} = 10k\Omega$ and $t_p = 932\mu s$. In green is highlighted the needed energy, while in red the wasted contribution, marking the values at the mean switching time t_m, and at distribution tails t_σ and t_σ.

3.3.3 *SPICE simulation in CVS and CQS*

We implemented the equivalent RRAM circuit model, presented in subsection 3.3.1, for simulation of a set event in CVS and CQS, in order to compare and benchmark the two approaches. The simulation parameters were extracted from electrical



(a) TEM picture of a $TiN/HfO_2/Ti-TiN$ sample in 1T1R cell configuration.

(b) Energy consumption during a set process versus the switching time t_s, calculated according to (11) at a pulse amplitude of 1V.

Figure 42 – Samples tested in CVS, and relative switching times and energy distribution

Parameter	$V_{th}(V)$	$V_{set}(V)$	$R_{HRS}(\Omega)$	$R_{LRS}(\Omega)$	$t_p(s)$	$E_{th,99.5}(J)$	$E_{th,m}(J)$
Value	0.5	1	100k	10k	1m	9.3n	230p

Table 4 – Parameters adopted for set energy consumption simulation, when using the equivalent circuit proposed in Fogure 41.

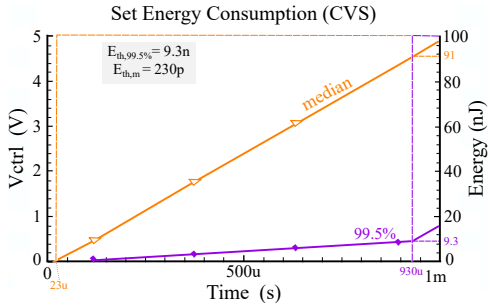
tests according to how illustrated in the previous subsection 3.3.2. Table 4 lists the values adopted.

Where $E_{th,m}$ and $E_{th,\sigma}$ represent, respectively, the set switching energy at the median and σ of the distribution. In CVS mode, the programming source is a constant voltage of amplitude 1V and time duration $t_p = t_{p,99.5\%} = 940\mu s$. The energy dissipated by the sample over the writing phase can be written as:

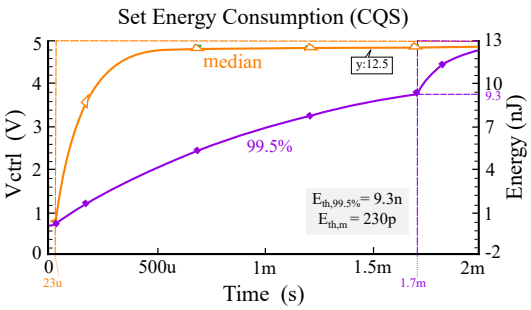
$$E(t)_{CVS} = \begin{cases} \frac{V_{set}^2}{R_{HRS}} t & \text{if } t < t_s \\ \frac{V_{set}^2}{R_{LRS}} t & \text{if } t_s \leqslant t \leqslant t_p \end{cases} \tag{17}$$

Therefore, the energy curve is a linear ramp, whose inclination is equal to V_{set}^2/R_{HRS} before the set occurs, and to V_{set}^2/R_{LRS} afterwards. Typically $R_{HRS} \gg R_{LRS}$, so that the rate of energy consumption after the set has taken place, R_{HRS}/R_{LRS}, is significantly higher than in HRS. Figure 43a reports a time transient analysis, computed from $t = 0$ to $t = 1ms$, illustrating the energy dissipation for samples at the median μ and at the 99.5 percentile of the distribution. It can be seen that, due to

the large dispersion of the switching time (energy), the consumed energy at the median leads to large energy waste.



(a) Energy consumption in CVS mode. The trend is a straight line, whose inclination increases after the set reaches the LRS state, in response to the higher flowing current into the cell.



(b) Energy consumption in CQS mode. Despite the set time dispersion, the final energy equals to 12.5pJ, as it is bounded by the amount initially stored inside the programming capacitor.

Figure 43 – Spice simulation of a set event, adopting the equivalent circuit shown in Figure 41. In solid colors, the energy consumption for a cell switching at the mean set time (23µs) and at the 99.5^{th} percentile.

On the other hand, when a capacitor is used for the set, the energy that can be delivered to the sample is bounded, as expressed by equation 8. Therefore, it is required that the programming capacitance is sufficiently large, in order to guarantee a successful operation. Given 8, and posing as boundary condition a minimum residual voltage over C, $V(\infty) = V_{th}$, the energy available to set a device can be expressed by:

$$\Delta E = \frac{1}{2} C (V_{set}^2 - V_{th}^2) \tag{18}$$

which can be rearranged to derive a condition on the capacitor size:

$$C_{min} = \frac{2E_{th,SR_{99.5\%}}}{V_{set}^2 - V_{th}^2} \tag{19}$$

where C_{min} is the minimum required capacitance. Posing $E_{th,SR_{99.5\%}} = 9.3nJ$, $V_{set} = 1V$ and $V_{th} = 0.5V$, it results:

$$C_{min} = 24nF \tag{20}$$

Figure 43b reports a Spice simulation of a set process in CQS mode. A programming capacitor of 25nF, precharged to 1V, is used to set a memory cell. Similarly to the previous case, samples switching at $E_{th,m}$, $E_{th,m}$ and $E_{th,\sigma}$ are considered. It is straightforward to see that, independently to the required threshold energy, the energy curves at the end of set process are asymptotic to the energy stored inside the programming capacitor (12.5nJ, assuming a full discharge). Therefore, although some overprogramming still occurs, for those cell that require less than

12.5nJ to transit to their LRS, the energy waste is strongly reduced with respect to the CVS case. Moreover, the programming energy variability is nulled, as the delivered energy is fixed.

3.4 ENERGY COST COMPARISON

The programming efficiency, η, for a set process, can be readily derived from:

$$\eta = \frac{E_{th}(t_{set})}{E_{tot}} = \frac{E_{th}(t_{set})}{E_w + E_{th}(t_{set})} \tag{21}$$

where $E_{th}(t_{set})$ is the energy needed to set a cell whose switching time is t_{set}, and E_{tot} the total energy consumed. Declining (21) for the CVS case, and substituting E_{th} and E_w with expressions (11) and (14), it holds:

$$\eta(CVS) = \left[\left(\frac{t_p}{t_{set}} - 1\right)\frac{R_{HRS}}{R_{LRS}} + 1\right]^{-1}$$

for the CQS case, expressing E_{tot} with (8), results:

$$\eta(CQS) = \frac{E_{set}}{E_c} = \frac{2t_{set}}{R_{HRS}C} \tag{22}$$

Figure 44 shows a plot of the calculated pgramming efficiencies. It can be seen that in the CQS case, the efficiency is able to raise above 50%, as the losses to charge the programming capacitor are not taken into consideration. Such result can be realistic, as observed in Section 3.2,if the programming capacitor is charged with an optimized charging process.

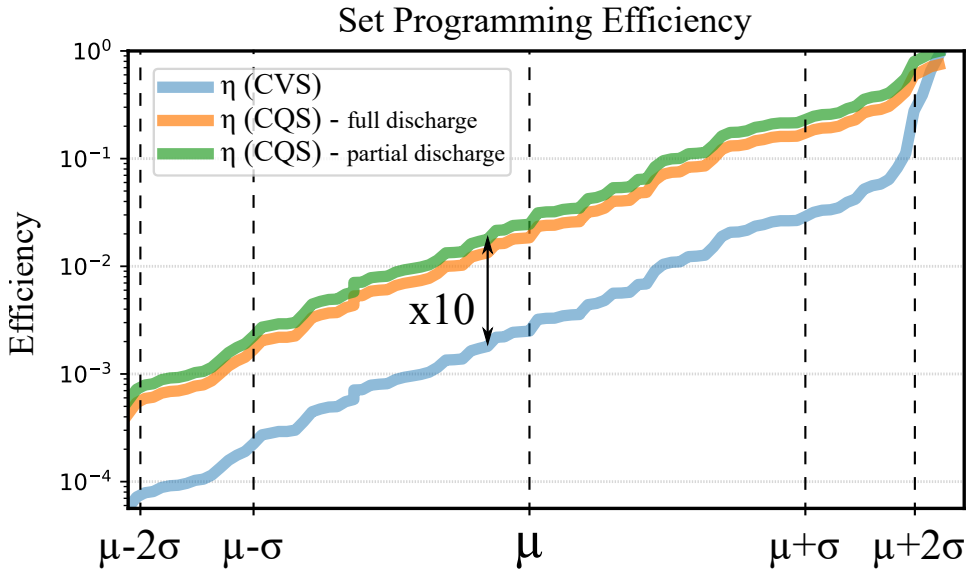


Figure 44 – Set efficiency comparison between CVS (blue) and CQS (orange and green).

The efficiencies look, overall, low: this result is however consistent with the large switching time variability observed, and the high process success rate demanded.

Moreover, the low voltage amplitude used to set the samples contributes to relatively long switching times at the upper tail of the distribution (this aspect is further commented in Section 3.6.1). Nonetheless, the proposed CQS capacitive switching shows a clear improvement over the CVS SOA approach, which is up to ~10x until the upper tail of the distribution. When the programming capacitor is fully discharged onto the load (orange curve), the CVS efficiency becomes better than the CQS after the 99 percentile, due to the fact that the energy delivered by the programming capacitor (12.5nJ) is greater than the maximum required (9.3nJ). If the discharge is stopped at $V_{th} = 0.5V$ (green curve), the efficiency can be improved, and is always superior to the SOA method.

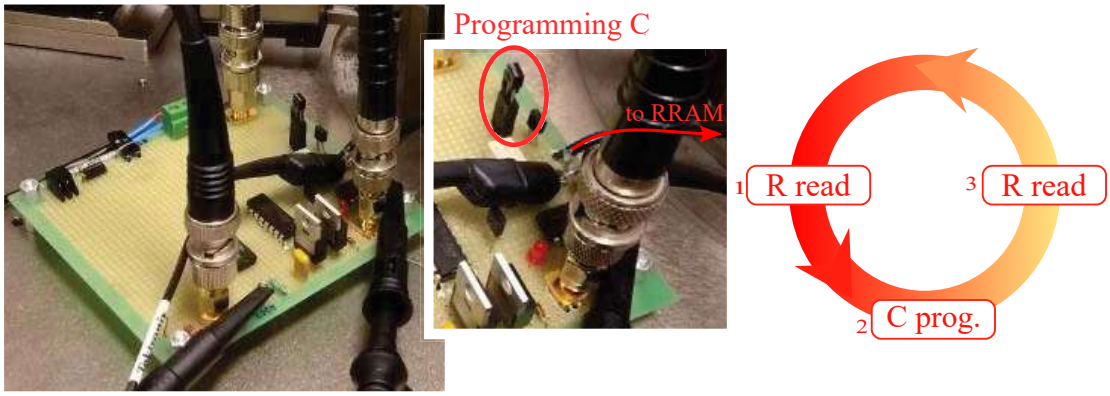
CQS PROGRAMMING TIME EXTENSION Despite the remarkable gain in energy efficiency, a major drawback is that the programming time in CQS ($t_{p,CQS} = 2ms$) is larger than the one required in CVS mode ($t_{p,CVS} = 2ms$). This extension derives from the fact that a charged capacitor requires a longer time to deliver the same amount of energy, compared to a constant voltage: as the discharge takes place, the voltage over the capacitor drops, delivering energy at a slower rate. Consequently, the time gap between the CVS and CQS case becomes greater the later the cell switches, the worst-case scenario being when the cell sets at the minimum residual voltage, V_{th}. The maximum switching time can thus be expressed by:

$$t_{s,max} = -R_{HRS,max} C_p \ln \frac{V_{th}}{V_{set}} \tag{23}$$

where $C_p = 25nF$ is the programming capacitor, and $R_{HRS,max}$ the maximum HRS resistance value. Given our framework, where no statistical resistance variation is considered, so that $R_{HRS,max} = \overline{R_{HRS}}$, $t_{s,max}$ is equal to 1.73ms. However, expression (23) properly reflects a practical case only when $R_{HRS,max}$ is replaced with its actual value. Although the dilation of programming time looks problematic, this side-effect can be effectively limited by increasing the programming capacitor initial voltage, as further explained in Section 3.6.

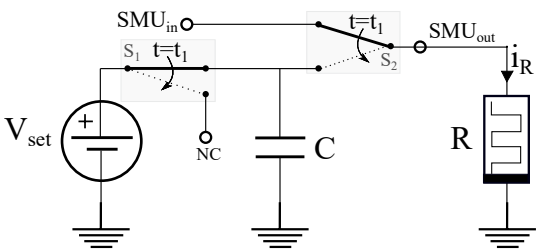
3.5 EXPERIMENTAL BOARD FOR PROOF OF CONCEPT

In order to deliver proof of concept of the newly proposed CQS programming approach, the circuit board illustrated in Figure 45 was developed. Figure 45a shows a picture of the realized circuit, the inset highlighting the programming capacitor, which is a discrete component inserted into the predisposed female headers. Such setup allowed to handily test a wide range of different capacitance values. The board has three voltage inputs, which are supplied by Parameter Analyzer HP4155/56: SMU_{in}, VSU and $TRIG_{in}$. SMU_{in} is used to read the device-under-test resistance, while VSU is a constant voltage source used to charge the programming capacitor to its initial voltage, V_{set}; $TRIG_{in}$ is a train of pulses that is acquired by the circuit board and turned into a step voltage in order to synchronize the board with the Parameter Analyser. An equivalent circuit illustrating the board operation is shown in Figure 45c, while a full schematic is reported in Figure 45d.

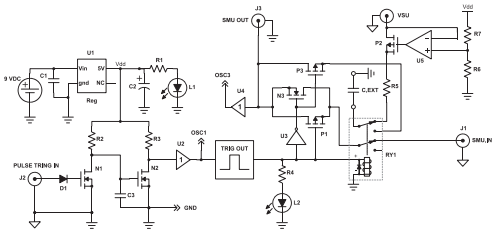


(a) Discrete-component circuit board. The zoomed inset highlights the programming capacitor.

(b) Operational sequence



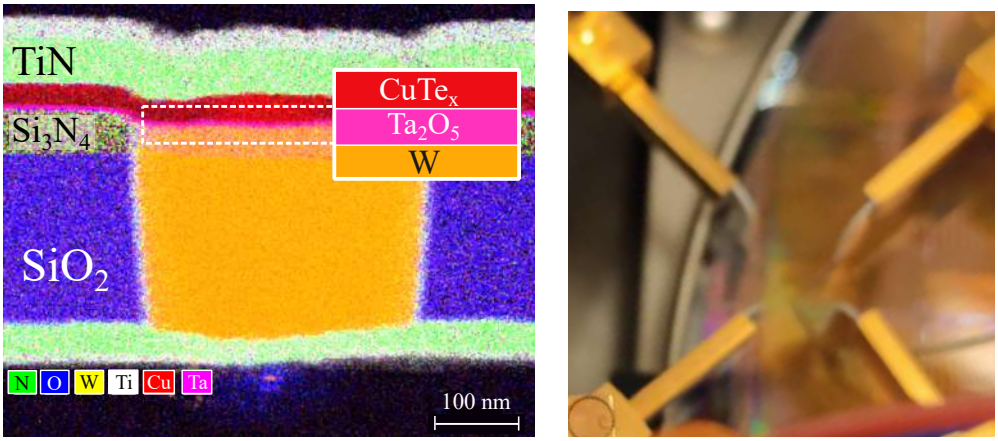
(c) Equivalent circuit



(d) Full schematic

Figure 45 – Circuit board to demonstrate proof of concept of the proposed capacitor-based RRAM programming

At $t = 0$, SMU_{in} is sourcing a non-destructive voltage of 100mV to read the DUT resistance; 10 samples of current are collected and averaged out to have a reliable read. Contemporarily, the VSU input supplies V_{set} to charge the programming capacitor C. Successively, at $t = t_1$, $TRIG_{in}$ signal changes logic state and commutes relay S_1 and S_2, so that C is isolated from the supply and coupled onto the RRAM cell. After the programming capacitor is fully discharged, the resistance of the sample is read again to verify the success of the set operation. Figure 45b reports the test operational sequence. Multiple cycles can be performed if the resistance after the programming operation is not low enough, until a satisfying value is achieved. The connection from the board output, SMU_{out}, to the DUT is performed by means of a short cable terminating with micro SMA connectors, in order to minimize the parasitic capacitance in parallel to the sample. Figure 46a shows a TEM picture of the tested RRAM cells, of stack $CuTe_x/Ta_2O_5/W$ in 1R configuration. In Figure 46b, a photo of a device on wafer, the top electrode being connected to the board output SMU_{out}, while the bottom is grounded.
The whole test setup is shown in Figure 47. A computer running a dedicated Phyton script remote-controls the Paramenter Analyser, in order to supply bias to the board and retrieve measurement data. Thanks to the presented configuration, the whole test sequence was automatized in order to allow extensive testing without the need for manual intervention.
Table 5 presents some preliminary results. Test parameters were the programming capacitance, which ranged from a few nF to 0 (which corresponded to no capacitor



(a) TEM picture showing the sample stack (b) picture of a DUT on wafer

Figure 46 – Characterized 1R RRAM device.

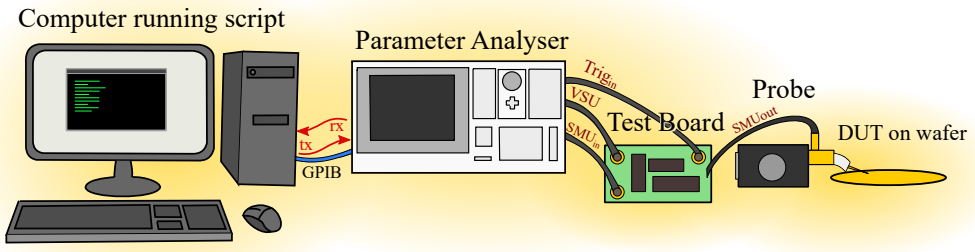


Figure 47 – Full test setup for conducting CQS programming, employing the designed test board.

plugged inside the board headers, so that only parasitic capacitance was present), and the charging voltage V_{set}, which was set to either 3 or 5V. For the tested transistor-less cells, the set operation corresponded to the forming operation, so that relatively high charging voltages were required. For each testing condition, i.e. at a given capacitance and charging voltage, 3 samples were tested, in order to conclude if the process was successful or not. The intrinsic limitations of a discrete component, self-soldered board do not allow for a precise quantification of the energy delivered to the sample during the so-performed CQS switching. However, the collected results follow a logical trend, as a set process was rarely achieved when the capacitance value was decreased below 330 pF, and never in its absence. For this reason, the built setup still serves well as proof of concept to demonstrate that a capacitor-based set operation is achievable, and inspired further work on the implementation of an integrated-circuit solution, which would bring much higher control and precision.

Capacitance (F)	Set Voltage (V)	Successful set
$> 1n$	5	yes
$1n$	5	yes
700p	$5-3$	yes
470p	$5-3$	yes
330p	$5-3$	yes
$< 330p$	$5-3$	no

Table 5 – Results with various capacitance values at different charging voltage.

3.6 DESIGN CONSIDERATIONS AND IMPORTANT TRADE-OFFS

The analysis exposed in the previous sections allowed to conclude that the proposed capacitor-based programming method can effectively raise the programming efficiency of a set operation. However, the process efficiencies derived in Section 3.4 look relatively low: $\eta(CVS_{\mu+\sigma}) = 2\%$ and $\eta(CQS_{\mu+\sigma}) = 10\%$. As previously mentioned, this result should not surprise when considering the testing conditions, which led to a broad dispersion of the switching times, and considerable overprogramming to ensure a high success rate. The high energy demanded by the slowest switching cells translated in a relatively big programming capacitor (25nF), which is too large for realistic on-chip integration. A narrower dispersion of the switching times, and an overall faster set process, would help raise the process efficiency while reducing the energy cost. As a result, the proposed method would represent a realistic, integrable, alternative to the SOA.
The work conducted by G. Sassine et al. [116] delivers useful insights on different programming conditions, obtained by varying the set time, current and voltage. The tested technology is the same that was adopted in our evaluation, which allowed us to draw further considerations on our study based on their conclusions. Figure 48 summarizes their findings: it shows the Window Margin (WM) at 2σ of the distribution, calculated after 10^5 cycles in endurance tests, versus the set energy [1]. Their results point in the direction that increasing the set current, I_{cc}, has a more beneficial impact on the window margin, with respect to extending the set time t_p. Moreover, by increasing the set voltage, the energy cost could be reduced to only 10pJ. Such decrease in energy looks outstanding, and, as the following Subsection illustrates, it is motivated by the fact that there is an inverse exponential relationship between the programming voltage and the set switching time. Therefore, as further expanded in Subsection 3.6.3, it is envisage-able to integrate the proposed CQS technique on a chip, where the programming capacitor is dramatically downsized to a few pF.

1. The set energy was estimated at the first order by $E = t_p V_{set} I_{cc}$, where t_p, V_{set} and I_{cc} are, respectively, the programming time, voltage, and current.

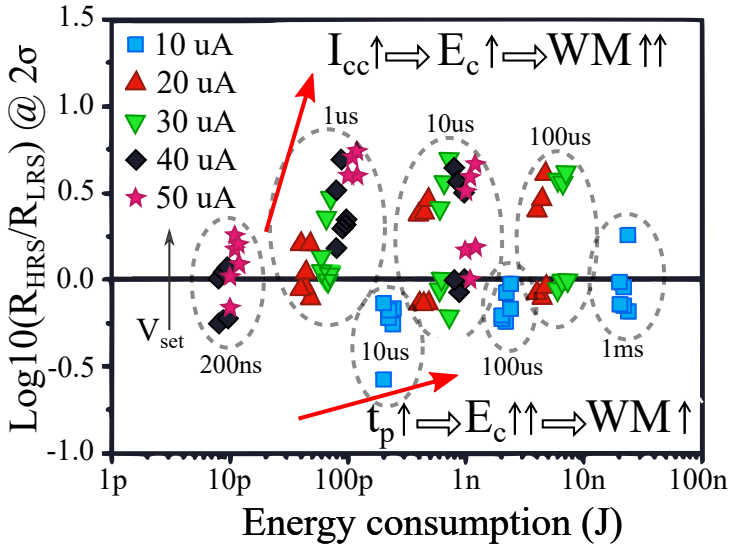


Figure 48 – Window margin (at 2σ) extracted from endurance CVS tests, at various programming conditions. Adapted from [116]

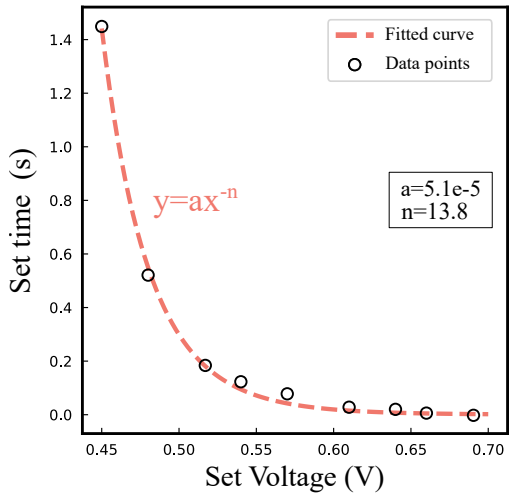
3.6.1 *Impact of the programming voltage onto the set switching time*

The set process in RRAM has been recognized as a soft-breakdown phenomenon [117, 118] and, as such, the memory switching time can be expressed by the time-to-breakdown T_{BD}. The literature highlights a strong voltage-dependence of the set time[2] to the set voltage, where an increase in potential is responsible of accelerating the phenomenon according to a power-law [117]:

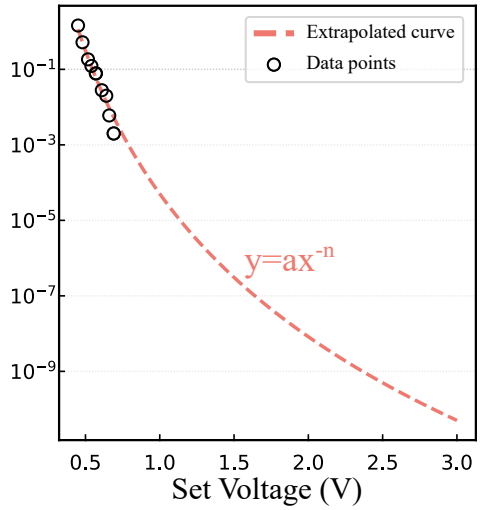
$$T_{set} \propto aV_{set}^{-n} \tag{24}$$

where V_{set} is the set voltage, corresponding to the voltage drop across the oxide, and a and n are constants.
In order to derive the actual voltage-time relationship for our samples, expression 24 has been used to fit the experimental data[3] reported by Sassine *et al.* [116]. The resulting curve is shown in Figure 49a along with the fitting parameters. Figure 49b reports an extrapolation for higher voltages: it can be seen that the switching time drops to the nanosecond range when the programming voltage is increased beyond 2V.
Different research groups have published on the set time shrinkage through the raise of the programming voltage amplitude, with the fastest experimental measures being as low as tens of picoseconds [119, 120].
Figure 49c shows the extrapolated energy amount at a given set voltage and time: it can be seen that a faster switching process allows to greatly reduce the energy cost. As the following Subsection illustrates, this allows to implement a relatively small capacitor to integrate the CQS approach on chip.

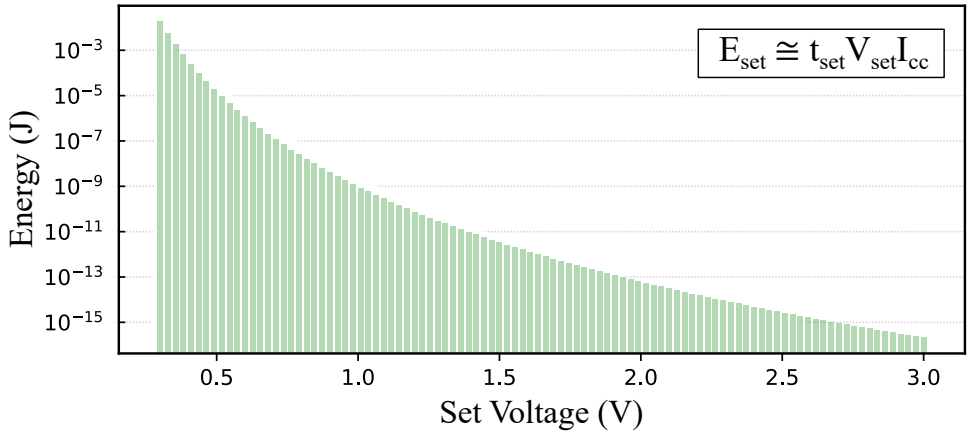
2. Set time denotes the time interval required for a sample in HRS to transit to its LRS.
3. The set times considered are those located at the upper whisker of the distribution (2σ).



(a) Set switching time versus the set voltage. The measurement data points were fitted with the dotted curve.



(b) Extrapolated fit for increasing set voltages.



(c) Extrapolated set energy consumption versus the set voltage. The current compliance value adopted was $I_{cc} = 100\mu A$.The energy cost drops as the set voltage is increased.

Figure 49 – Extrapolation of technological parameters for the RRAM technology under analysis, and projection of the energy cost at increased set voltages.

3.6.2 *A dynamic RRAM model*

When a charged capacitor is implemented to program a memory cell, its voltage, $V_C(t)$, is time-varying over the set time. Therefore, a RRAM model that takes into account the time evolution of the programming voltage is required, in order to reflect the behaviour of the cell more realistically. We propose a Verilog-A model, which implements the operational flow illustrated in Figure 50: at each time interval, the switching conditions are checked, and the voltage over the memory is replaced by its RMS (Root Mean Square) value. At the top view, the model has one input: V_{mem}, the instantaneous voltage across the RRAM, and one output: Set, a flag equal to 0 when the memory is in HRS, and 1 when in LRS.

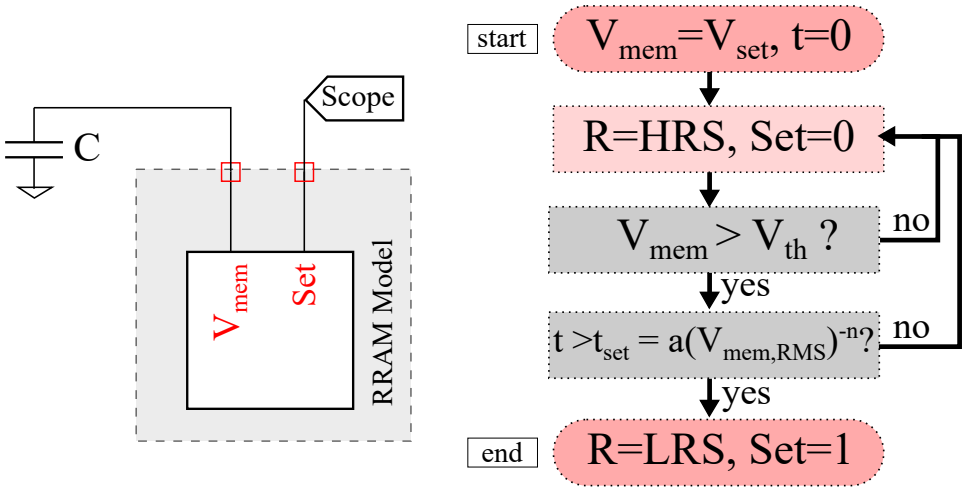


Figure 50 – Dynamic Verilog-A RRAM model, implemented to simulate a set event in the context of a varying voltage across the memory cell.

3.6.3 *CQS programming integration in a RRAM matrix*

As discussed in the previous Section, the increase of the programming voltage entails a strong energy cost drop, which in turn allows to highly downscale the programming capacitor. This section exposes some further evaluation on the integration of the proposed CQS technique in a RRAM matrix.
The memory array brings extra parasitic contributions, which have to be taken into account in order to properly size the programming capacitor. Figure 51 illustrates this concept, where element *matrix par* is added to the basic CQS topology. The inset shows the equivalent circuit which models the losses brought by the array non-idealities: C_{par} stands for the parasitic capacitance at the bitline, R_{par} the resistance of the metal lines and switches/multiplexers, and R_{leak} the overall leakage to ground.

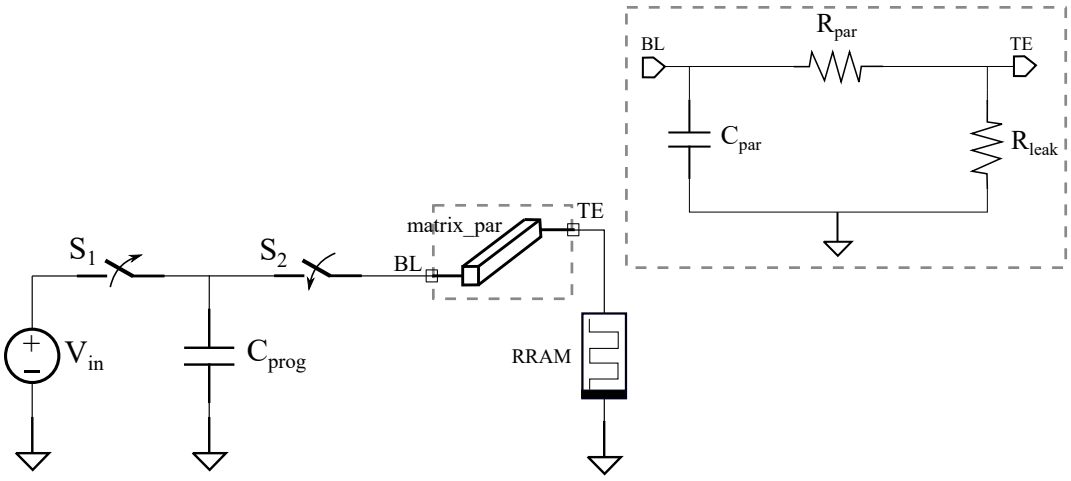


Figure 51 – Equivalent circuit including the parasitic contributions brought by the memory array.

The parasitic contributions should be quantified according to technology specifications and post-layout simulation, so that the circuit shown in figure 51 can be used for reliable simulations. Ultimately, the energy that is delivered to the RRAM load is expressed by:

$$E = \frac{1}{2}((V(0)\alpha)^2 - V_{th}^2) \cdot (C_{prog} + C_{par})(1-\beta) \cdot \gamma \tag{25}$$

where $V(0)$ is the charging voltage of programming capacitor C_{prog}, V_{th} the minimum voltage required to trigger the set event, and α, β, γ adimensional constants that quantify losses, respectively:

$$\begin{aligned}
\alpha &= \frac{C_{prog}}{C_{prog} + C_{par}} && = \text{ charge sharing losses} \\
\beta &= \frac{R_{par}}{R_{par} + R_{HRS} // R_{leak}} && = \text{ series resistance losses} \\
\gamma &= \frac{R_{leak}}{R_{leak} + R_{HRS}} && = \text{ leakage losses}
\end{aligned}$$

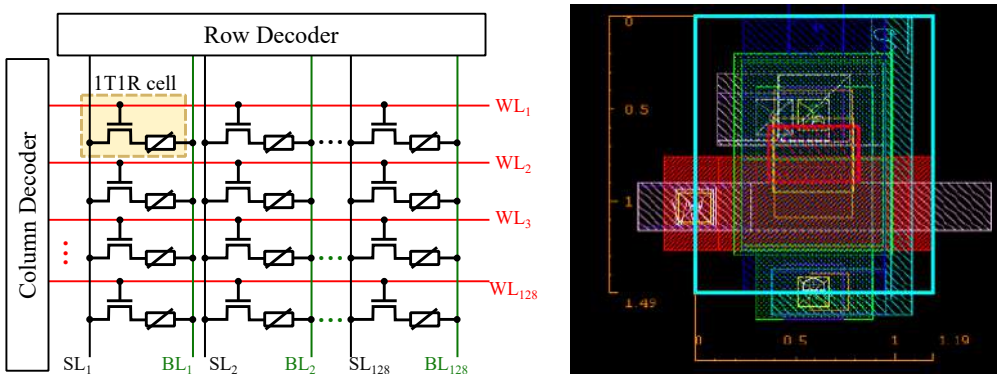
The expression for the minimum required capacitance, $C_{prog,m}$, can be written from Equation (25) by substituting E with E_{th}, the threshold energy needed for programming a RRAM cell, at the desired success rate. Solving for C_{prog}:

$$C_{prog,m} = \frac{\sqrt{E_{th}^2 - V(0)^2 C_{par}(1-\beta)\gamma[2E_{th}(C_{par}+1) + V_{th}^2 C_{par}(1-\beta)\gamma^2(1-C_{par})^2]}}{(1-\beta)\gamma(C_{par}-1)[V(0)^2 - V_{th}^2]} \tag{26}$$

by adequately tuning the boundary conditions, Equation (26) can be used for both a set and reset operation: E_{th} being equal to either $E_{th,set}$ or $E_{th,reset}$ and V_{th} to $V_{th,set}$ or $V_{th,reset}$, respectively. However, as the energy required for reset is generally larger (by typically an order of magnitude) than a set, the capacitor results larger in the former case.

3.6.4 *CQS process integration in 16kb array*

We proceeded our evaluation by considering the specific case of a 16-kb RRAM matrix, whose schematic is shown in Figure 52a. The elementary cell configuration, also called *bitcell*, consists in a 1T-1R structure (highlighted in figure). The full array comprises of 128x128 rows and columns, which are addressed by the respective decoders.
Figure 52b shows the bitcell layout. The access transistor is integrated with 130nm CMOS technology, while the RRAM cell, fabricated at a later BEOL stage, is placed between the top $M4 - M5$ metal layers. In order to extract the parasitic RC contributions brought by the memory array, post-layout simulation on the matrix layout was performed.
We then proceeded to extract the threshold values to achieve a set operation for our RRAM technology; E_{th} and V_{th} were chosen so that a window margin of 10 could be achieved at 2σ of the distribution, according to the experimental evidence



(a) 16kb matrix schematic.

(b) Bitcell layout view, showing a RRAM on top of the access transistor.

Figure 52 – Layout of a rram matrix structure, showing three 16kb arrays side by side.

Parameter	$E_{th,set}(J)$	$V_{th,set}(V)$	$C_{par}(F)$	$R_{par}(\Omega)$	$R_{leak}(\Omega)$	$R_{HRS}(\Omega)$	$R_{LRS}(\Omega)$
Value	10p	1	53f	2k	7G	100k	10k

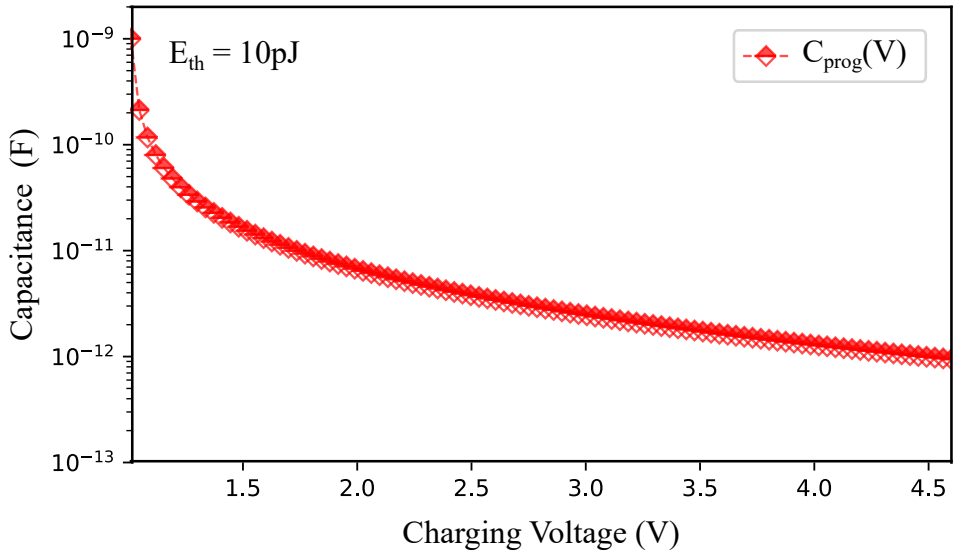
Table 6 – Extracted parameters for in-matrix simulation.

gathered by Sassine *et al.* [116]. The parameters used for simulation are reported in Table 6.

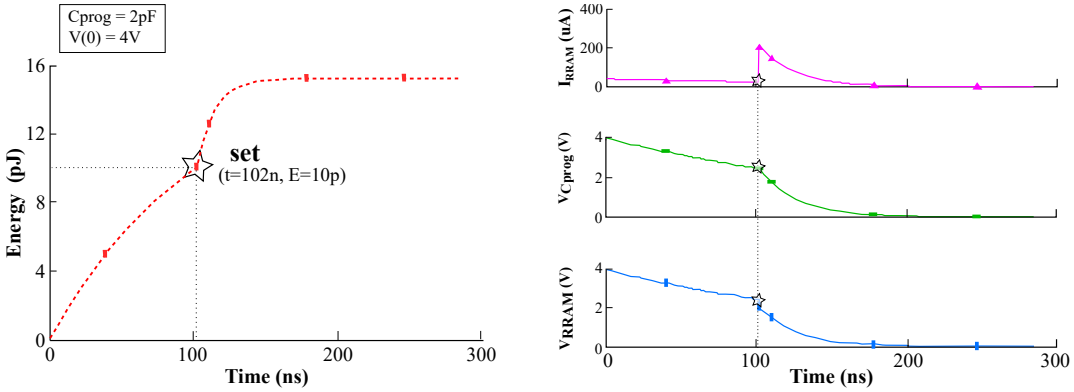
Next, the programming capacitor can be sized using Equation (26). The results are shown in Figure 53a, which plots the programming capacitance versus the charging voltage, ranged in the interval $]1, 4.6]V$. Noticeably, the higher the initial voltage, the smaller the required capacitance; this result should not look surprising, as the energy stored in a capacitor scales quadratically with the voltage and linearly with the capacitance.

Moreover, although the value of $E_{th,set} = 10pJ$ has been adopted to reflect experimental observation [121], Figure 49c suggests that the set energy should decrease as the set voltage is increased: consequently, a lower capacitance range could be targeted when the initial voltage is increased above 1V. This observation opens the possibility of either downsizing the programming capacitor, or using the parasitic bit line capacitance alone for the set of a memory cell, reducing the area cost of the proposed approach.

Figure 53b shows Spice simulations of the switching process, performed using the dynamic RRAM model presented in Subsection 3.6.2, and the matrix equivalent circuit of Figure 51. The values adopted are those reported reported in Table 6. The charging voltage, $V(0)$, was set to 4V, so that the minimum required capacitance for a set process resulted 1.3pF. A programming capacitor of 2pF was thus implemented.



(a) Programming (set) capacitance versus the charging voltage. The value is estimated taking into account the parasitic losses, using Equation (26).



(b) Spice simulation waveforms showing a set event in a 16kb RRAM matrix, where the parasitics limit the energy delivered to the memory cell. A capacitor of 2pF, charged at 4V, is able to program a cell with high success rate (figure to replace for better clarity).

Figure 53

3.7 A DUAL APPROACH: L-BASED PROGRAMMING

In this chapter, we have discussed the concept of programming a RRAM cell by means of a (pre) charged capacitor, due to its inherent ability of controlling the programming energy by the amount of stored charge. It thus follows natural at this point to consider enlarging the concept to its dual component: the inductor, where the energy can similarly be stored inside its magnetic field. Figure 54 shows the procedural sequence, where a charged inductor is charged and later switched onto a resistive load. Analogously to the capacitor case, if enough energy is delivered to the memory element, while a minimum voltage drops over the cell, the resistance transition can be triggered.

The inductor-based programming thus appears similar to the CQS method. Nevertheless, there are important differences and challenges to its actual implementation: for example, integrated inductors suffer from low quality factor and restricted

value range, so that a off-chip component (with all its downsides) should be envisaged in this case. The following Section exposes in further detail the main features of RRAM programming by means of a charged inductor, along with a final comparison between C and L-based programming.

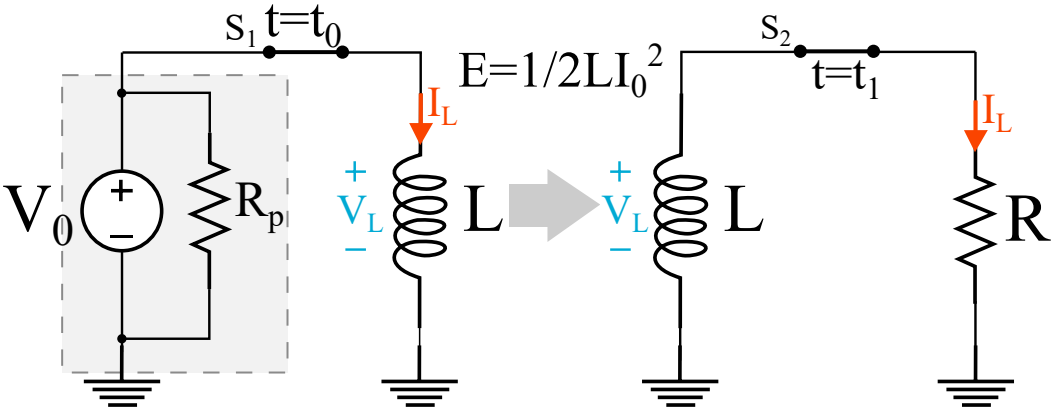


Figure 54 – Inductor-based programming sequence, comprising of a charging and discharging step. The load is simplified by a constant resistor

3.7.1 *Overview of L-based writing process*

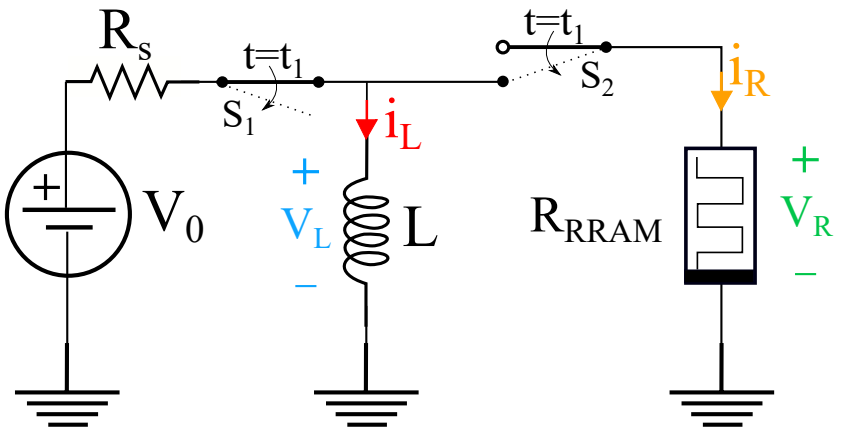
As opposed to a capacitor, an inductor tends to keep constant the amount of current flowing through its terminals. Any change in current is "contrasted" by a rise in voltage, according to the inductor's law:

$$V_L = L\frac{di_L}{dt} \tag{27}$$

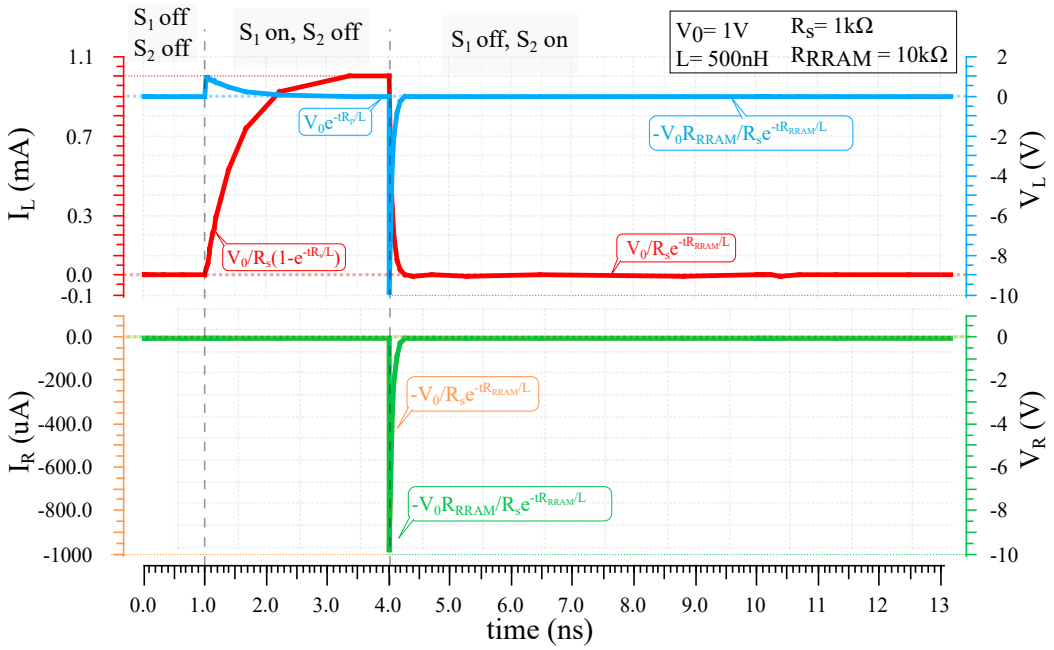
where L is the inductance and di_L/dt, the rate of current change. Equation (27) predicts that a fast change in i_L will cause V_L to rise, potentially to dangerously high values. Such situation could be encountered when the system is powered up, or the switches are commuted [4]. Consequently, precautions need to be taken in order to limit the voltage amplitude over the programming inductor. At the charging phase, this problem can be avoided by charging the inductor with a voltage source; moreover, this solution allows to obtain higher charging efficiency [5]. Figure 55a illustrates an updated circuit, whose Spice simulation results are reported in Figure 55b. Although the closing of switch S_1 causes an abrupt current change in the initially discharged inductor, the voltage across L is clamped by that of the source, V_0. The inductor's current saturates to the charging current $I_0 = V_0/R_s$, where R_s is the series resistance.
The critical transition takes place at the start of the discharging phase, at $t = t_1$, when the inductor is disconnected from the charging source and coupled onto the load. As the inductor tends to maintain the same amount of current in open

4. This is the case when transistors are used to implement the switches, with advanced technological nodes having breakdown voltages in the range of few volts.

5. By charging with a voltage source, a current ramp develops across the inductor, so that the heat losses can be minimized according to the principle of adiabatic charging illustrated in Section 3.2.



(a) L-programming circuit, where the switches commute sequentially in order to first charge the writing inductor L and later tranfer energy to the RRAM load.



(b) Simulation waveforms.

Figure 55 – Spice simulation of L-based RRAM programming, where the RRAM is simulated as a constant load.

circuit condition, V_L peaks uncontrollably. If S_1 is opened shortly after closing S_2, in order to ensure that the charged inductor is never unconnected, the voltage that develops will depend on the memory's resistance (R_{RRAM}), namely:

$$V_L(t_1) = I_0 R_{RRAM} = \frac{V_0}{R_s} R_{RRAM} \tag{28}$$

In the case of a current decrease, if $R_{RRAM} > R_s$, V_L will have negative sign. An higher voltage will be obtained when the memory cell is in HRS, i.e. during a set operation: in this case $V_L(t_1) = I_0 R_{HRS}$. Considering the high values and wide dispersion of the HRS resistance (the mean $\overline{R_{HRS}} = 100k\Omega$ for our technology) it is straightforward to see that V_L can easily become very high.

By looking at Equation (28), one might consider minimizing the charging current in order to limit the voltage overshoot; however, this approach might still be insufficient against the wide HRS distribution, and also lowers the chaging efficiency. Most importantly, the energy stored inside the inductor, E, decreases (quadratically) when lowering the charging current:

$$\begin{aligned} E_L &= \frac{1}{2} L I_0^2 \\ L &= \frac{2E_L}{I_0^2} \end{aligned} \tag{29}$$

where L is the inductance in Henry and I_0 the charging current amplitude. It thus follows that more advanced design solutions have to be considered in order for the energy to be sufficiently high to perform the programming operation, whilst the voltage across the inductor is kept to a reasonably low value.

3.8 A PERSPECITIVE APPROACH: COMBINATION OF CHARGED CAPACITOR AND INDUCTOR PROGRAMMING

As explained in the previous Section, a charged inductor represents a dual approach to the CQS method illustrated in this chapter.
A programming capacitor was introduced in order to limit the energy waste during a set operation, which, when performed with the standard CVS technique, is highly energy inefficient. The reset operation was not discussed: in fact, the resistance increase as the memory transits to its HRS self-limits the energy waste. Moreover, the energy cost for a reset operation is considerably higher than the set, and thus demands a prohibitively large capacitor for on-chip integration. It follows that the CQS technique is not particularly attractive for a reset process.
On the other hand, a higher current demanding, and lower resistance state, is a friendlier framework to a charged inductor, as it allows to reduce the size of the programming element (see Equation (29)) and the overvoltage hazard. For example, considering a reset energy of $E_{res} = 100pJ$, and a programming current of $2mA$, the size of the programming inductor can be estimated (using Equation (29)) around $L = 100nH$, which could be integrated off-chip.
Therefore, future work can be envisaged where an architecture implements a charged capacitor for set operations and a charged inductor for reset.

3.9 SUMMARY AND CONCLUSIONS

In this chapter, we proposed novel *energy-controlled* methods to perform the programming operation in RRAM. As opposed to the SOA technique, which implements a voltage pulse whose duration has to extend to the upper tail of the switching time distribution, leading to high energy waste, we propose to limit the energy dissipation by opportunely charging a programming capacitor.
We began our analysis by delivering a voltage and energy sensitive compact model, calibrated with experimental data on SOA RRAM technology. Succesively, we delivered proof of concept with a discrete-component board. By exploiting the

power-law relationship between the set switching voltage and time, we demonstrated that a conveniently small capacitance (2pF) can be obtained for on-chip integration, albeit care has to be taken to the parasitic losses when designing an in-matrix protocol.
As a final step we proposed a dual approach, centered on a charged inductor. Given that L can be decreased with the square law of the programming current, and that an inductor can generate large voltage overshoots in the presence of high impedance loads, we consider that a charged inductor might be more beneficial when used for a reset process, rather than a set. Therefore, we propose as prospective work an architecture where a charged capacitor is used for set operations, whereas a charged inductor for reset.
Table 7 summarizes the different programming approaches presented in this chapter, and reports their major advantages and disadvantages.

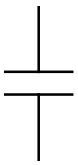
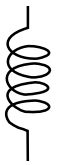
	Voltage pulse	**C-programming**	**L-programming**
Type			
E_{set}	$\frac{t_{set}V_{set}}{R_{HRS}}\left(\frac{t_p - t_{set}}{t_{set}} + \frac{R_{HRS}}{R_{LRS}}\right)$	$\frac{C}{2}(V_{set}^2 - V(\infty)^2)$	$\frac{L}{2}(I_{set}^2 - I(\infty)^2)$
Min. t_p	t_{set}	$R_{HRS}C \cdot \ln\left(\frac{V(t_{set})}{V(0)}\right)$ (neglecting parasitics)	$\frac{L}{R_{HRS}}\ln\left(\frac{V(t_{set})}{V(0)}\right)$ (neglecting parasitics)
Pros	○ Standard method. ○ No need to know E_{set} in advance. ○ Compatible with bot set and reset operations.	○ High programming energy efficiency. ○ Reduced electrical stress during programming. ○ No need of feedback control during programming (such as WT).	○ High programming energy efficiency. ○ Reduced electrical stress during programming. ○ Current control. ○ Programming time decrease: $t_p < t_{set}$
Cons	○ Low programming energy efficiency. ○ High electrical stress during programming. ○ Degraded endurance and variability. ○ Need of a series element to keep current within a compliance limit.	○ Increased area footprint to integrate C. ○ Programming time extension: $t_p > t_{set}$. ○ Possibly damaging current overshoot. ○ Not suitable for reset operation.	○Inductor off-chip ○ Danger of high voltage overshoot. ○ Not suitable for set process.

Table 7 – Summary Table comparing the different programming strategies proposed in this chapter.

3.10 CHAPITRE 3 - RÉSUMÉ EN FRANÇAIS

Dans ce chapitre, nous avons discuté du concept de programmation d'une cellule RRAM au moyen d'un condensateur (pré) chargé, en raison de sa capacité inhérente à contrôler l'énergie de programmation par la quantité de charge stockée. Cette approche permet de surmonter le gaspillage d'énergie élevé, ainsi que le stress électrique, qui affligent l'opération du set d'une mémoire résistive. Nous avons inspecté les principales caractéristiques de ce nouveau protocole d'écriture en utilisant un modèle compact sensible à la fois à l'énergie et à la tension. Utilisant des temps de commutation expérimentaux (à 1 V) pour une technologie RRAM basée sur HfO_2, nous avons pu prouver que l'efficacité de l'approche proposée est améliorée d'un facteur 10.

Cependant, il faut considérer que la charge d'un condensateur par une source de tension est une approche intrinsèquement inefficace, car la moitié de l'énergie stockée à l'intérieur du condensateur est perdue sous forme de chaleur. Pour éviter ce problème, un processus de charge adiabatique doit être effectué. Comme alternative, une source de courant doit être utilisée pour charger le condensateur. Ces solutions permettent essentiellement de restaurer la pleine efficacité de charge.

Nous avons ensuite procédé à la conception et à l'assemblage d'une carte à composants discrets, afin de fournir une preuve de concept. Nos essais expérimentaux ont confirmé la faisabilité de la méthode, bien que de nombreuses pertes supplémentaires soient présentes dans cette configuration, de sorte que le condensateur de programmation est considérablement plus grand au-dessus de 300pF) que les valeurs d'intégration réalisables.

Nous avons ensuite exploité la relation exponentielle entre la tension de programmation et le temps pour réduire la taille de le condensateur. Des rapports expérimentaux montrent que l'énergie blanche peut être réduite à 10pJ; nous avons donc utilisé cette valeur comme référence lors de l'extrapolation de la capacité requise. Compte tenu des contributions parasites présentes dans une matrice de memoire de 16kb, nous avons enfine pu réduire considérablement le condensateur de programmation à 2pF, en chargeant à 4V.

Enfin, nous avons prolongé notre étude en considérant une approche duale, où une inductance chargée est utilisée pour programmer une cellule mémoire résistive. En fait, l'énergie peut être stockée de la même manière à l'intérieur du champ magnétique, et correctement limitée par la taille de l'inductance et l'amplitude du courant de charge.
Cependant, comme les inducteurs s'opposent au changement instantané du courant circulant, il faut faire attention lorsqu'il s'agit d'événements de commutation rapides, car ils déclenchent des pointes de tension potentiellement dommageables.

L'opération de reset n'a pas été discutée : en fait, la La résistance augmente au fur et à mesure que la mémoire transite vers son HRS auto-limite le gaspillage d'énergie. De plus, le coût énergétique pour une opération de réinitialisation est

considérablement plus élevé que l'ensemble, et exige donc un condensateur d'une taille prohibitive pour l'intégration sur puce. Il suit- bas que la technique CQS n'est pas particulièrement attrayante pour un processus de réinitialisation. D'autre part, un état de demande de courant plus élevé et de résistance plus faible est un cadre plus convivial à une inductance chargée, car il permet de réduire la taille de la l'élément de programmation et le risque de surtension. Pour l'examen- ple, en considérant une énergie de réinitialisation de Eres = 100pJ, et un courant de programmation de 2mA, la taille de l'inductance de programmation peut être estimée autour de L = 100nH, qui pourrait être intégré hors puce. Par conséquent, des travaux futurs peuvent être envisagés où une architecture implémente un condensateur chargé pour les opérations de réglage et une inductance chargée pour la réinitialisation.

Un résumé final et une comparaison entre la norme et les approches basées sur C et L sont donnés dans la dernière section du chapitre. Différents défis et avantages sont présents selon la méthode. Néanmoins, une architecture qui combine toutes les approches afin d'exploiter leurs traits les plus utiles pourrait être envisagée : par exemple, un jeu de programmation des condensateurs et des inductances pour les opérations de mise en service et de réinitialisation respectivement.

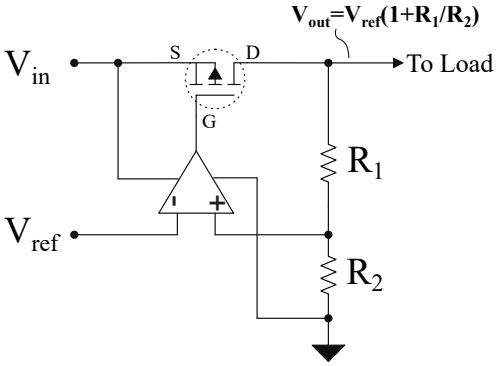
4

A SELF-LIMITING, PROGRAMMABLE CURRENT AND VOLTAGE SOURCE FOR RRAM WRITING

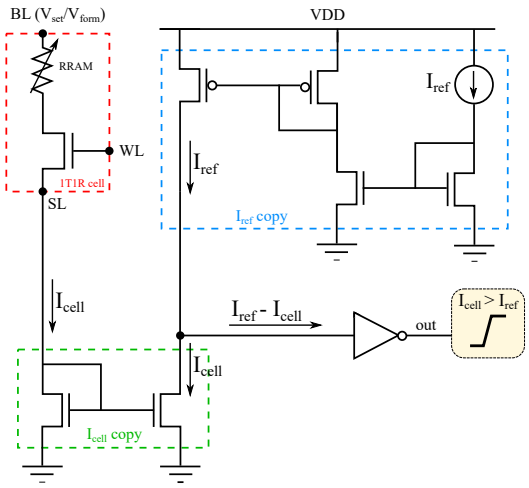
Chapter 3 discussed the standard technique for setting a RRAM cell, consisting in a voltage pulse of fixed duration, along with its major downsides: a high waste of energy and electrical stress for the target cell [60, 79, 121, 122]. In order to ease such side-effects, two new alternative programming methods were introduced, where the write operation is carried out by means of a charged capacitor, inductor, or a combination of the two. However, these approaches result in an enlarged area footprint, and possible overcurrent/overvoltage issues during resistance transitions.
This chapter delivers another novel design solution, where a current Digital to Analog Converter (DAC)-based circuit is implemented in order to both control the write voltage and current, and limit the set energy dissipation.
Section 4.1 gives an overview on the SOA write termination strategies implemented in RRAM arrays, while Section 4.2 presents our radically different proposition. Major design details are discussed in Section 4.3, and experimental results are shown and commented in Section 4.5.

4.1 WRITE TERMINATION CIRCUITS FOR RRAM PROGRAMMING

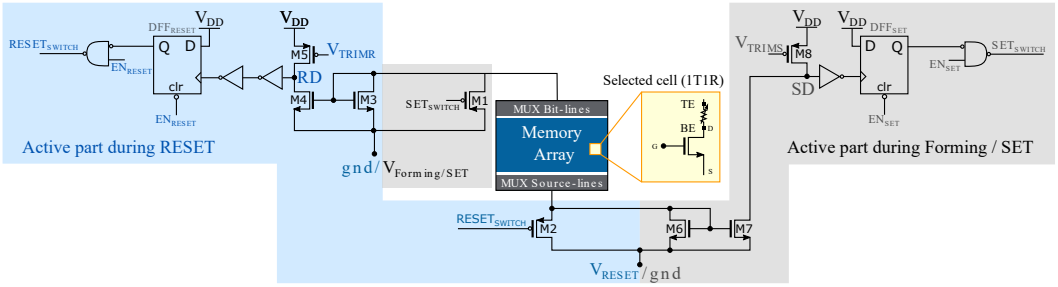
The term Write Termination (WT) stands for a class of circuits that interrupt the programming operation of a memory cell once the state transition is achieved, with the aim of reducing both the energy consumption and technological variability [121, 123–125]. Figure 56 shows a commonly implemented architecture in RRAM arrays, where the write voltage is generated by means of a Low-Dropout (LDO) regulator [125–128], while the current is detected/compared with current mirrors [123, 125].
The schematic of the voltage regulator is illustrated in Figure 56a: the output voltage, V_{out}, is controlled by means of feedback resistors R_1 and R_2, and the load current is delivered by the P-type transistor. Upon trimming of resistance ratio R_1/R_2, V_{out} can be varied, for example between the required amplitudes for a successful programming operation: V_{set}, V_{form} and V_{reset}. Subsequently, the selected cell Bit Line (BT) (Source Line, in the case of a reset) is coupled onto the LDO output. Figure 56b shows a common current-compliance detection schematic for a set/forming process, which is used to notify a successful transition to the LRS. At the un-buffered output, the current subtraction between the cell's current, I_{cell}, and the reference value, I_{ref}, takes place; as soon as $I_{cell} > I_{ref}$, the inverted output rises, delivering a trigger signal for the stop of the writing operation. A comprehensive schematic of a WT circuitry is reported in Figure 56c; the set/reset switches interrupt the cell biasing once the corresponding Flip-Flop (FF) output rises.
Such write termination approach has proved effective in both narrowing the resistance distribution and lowering the energy consumption [121, 123–125], which is



(a) Schematic of a **LDO!** (LDO!) regulator, used to generate the write voltage ($V_{set}/V_{form}/V_{reset}$) from the input bias V_{in}.



(b) Typical current detection circuit for a forming/set operation. The output rises when the cell current, I_{cell}, hits the reference value I_{ref} (compliance).



(c) Write termination circuit for forming/set/reset operations in a RRAM array. Adapted from [123].

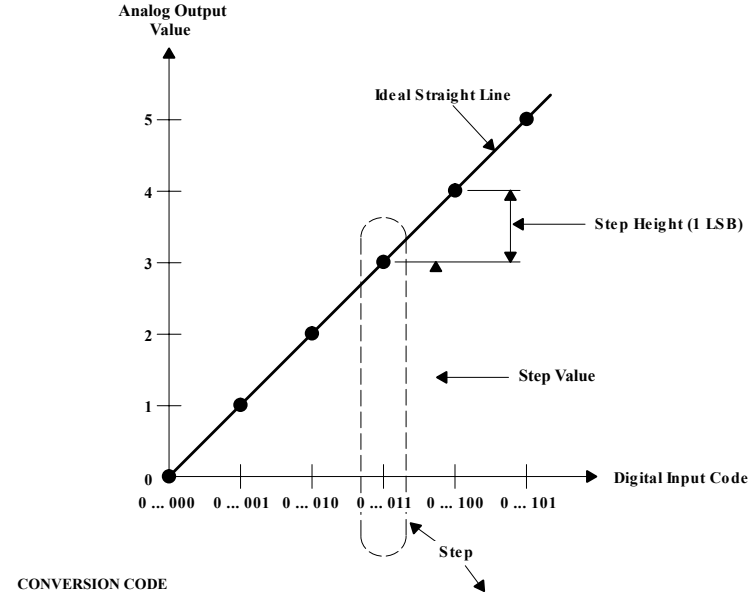
Figure 56 – Main circuital blocks of a common write termination architecture.

especially critical during a set process [79, 121]. However, memory performances are found to trade with power consumption, as [] reported that some *post-transition delay* is necessary to ensure satisfying memory-state retention and variability.
In contrast with the SOA architectures, our circuital solution is designed to be *self-terminating*, as voltage across the cell falls as soon as the resistance transition takes place. This way, the energy cost can be reduced while granting satisfactory memory operation.

4.2 PROPOSED CIRCUIT OVERVIEW

The motivation behind the proposed circuit is a well-controlled, current driven, set operation, which can effectively limit the current density inside a device transitioning to its LRS. In order to also regulate the voltage until the device resistance drops, our design converged into an integrated *current-switching DAC* architecture [129–132], whose ideal characteristic is shown in Figure 57. The transfer function consists in a straight line, connecting analog output values spaced by a Least Significative Bit (LSB). The input, a digital code, controls the analog output and sets its value among ground and the maximum analog voltage, in a way that resembles

a programmable potentiometer [132]. As a result, the system allows to tune the DAC output to the required set voltage.



CONVERSION CODE

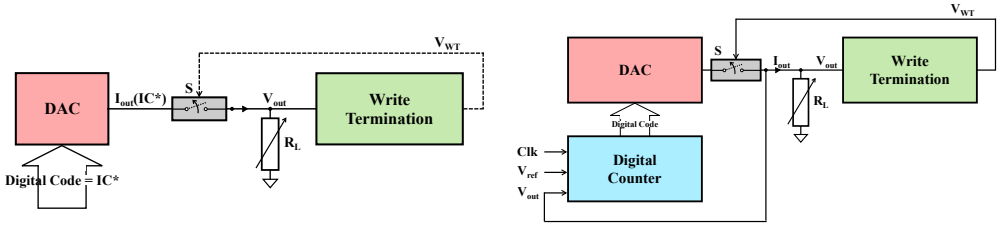
Digital Input Code	0 ... 000	0 ... 001	0 ... 010	0 ... 011	0 ... 100	0 ... 101
Analog Output Value	0	1	2	3	4	5

Figure 57 – Ideal transfer function of a DAC. Adapted from [132].

Figure 58 shows a system-level view of our design, which illustrates its dual working mechanism, as a programmable current or voltage source. In Figure 58a, a time-constant digital input sets the DAC output current, I_{out}, to a selected reference value I_{ref}. Given an opportune configuration of the input code, IC^*, the output current can be tuned, so that $I_{out}(IC^*) = I_{ref}$. Figure 58c shows qualitative time transient waveforms, which illustrate the behaviour of the circuit. If the write termination block is not active, the output current remains equal to the reference until the end of the programming time; else, the WT circuitry opens switch S when a resistance drop is detected. Figure 58b depicts the circuit operation when in voltage-source mode: the feedback loop driving the Digital Counter block regulates the output voltage, V_{out} to the reference V_{ref}. Qualitative time-transient curves are shown in Figure 58d: after settling time t_{settle}, V_{out} reaches V_{ref}, and is kept constant until the set event takes place. At t_{set}, the resistance drops causes V_{out} to fall; if the WT block is active, the voltage drop triggers the termination circuitry and cuts the bias across the load.

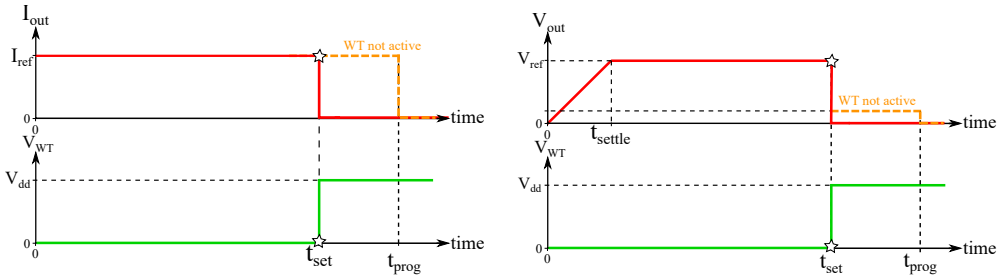
4.2.1 *Voltage Regulation*

Figure 59 illustrates in further details the circuit behaviour: it shows Spice simulation curves of V_{out} and I_{out} when the circuit is operated in voltage mode. More specifically, the circuit acts as a voltage source until the set event occurs. This



(a) Block diagram in current regulation (CCS) mode.

(b) Block diagram in voltage regulation mode.



(c) Time transient in CCS mode. The dotted line corresponds inactive WT block case. Write termination flag V_{WT} rises when a resistance drop is detected (at set time t_{set}), opening switch S.

(d) Time transient in voltage regulation mode. The dotted line corresponds inactive WT block case. At the set time, t_{set}, the load resistance drops and so does V_{out}. If WT is active, the bias is cut little after t_{set}, by opening switch S.

Figure 58 – Operational flow of the proposed circuit, in current and voltage drive mode.

is achieved by continuously comparing V_{out} to V_{ref}, and adjusting the Digital Counter's output consequently, so that:

$$V_{out}(t) = I_{out}(t) \cdot R_L = V_{ref}, \qquad \text{for } t > t_{settle} \tag{30}$$

Where t_{settle} is the system's settling time. The counter output is updated at each input clock rising edge, so that I_{out}, and hence V_{out}, are time-discretely regulated, taking the form of *stepped* waveforms. At start-up ($t = 0$), the counter initialises its count, and $V_{out} = I_{out} = 0$. After each clock's period (T_{clk}) the counting is incremented, until the output voltage tracks the reference. From this point on, the digital code oscillates between $count(t_{sett})$ and $count(t_{sett} + T_{clk})$. At t_{switch}, the set occurs, and R_L drops: the WT block detects a quick decrease of the output voltage and interrupts the bias across the memory cell.

4.3 CIRCUIT DESIGN

In this section, design details on the main operating blocks are presented: Section 4.3.1 and 4.3.2 cover the design of the current-switching DAC block, and illustrate its working principle. Section 4.3.3 focuses on the counter's architecture, while Section 4.4 presents the switching detection and write termination mechanisms.

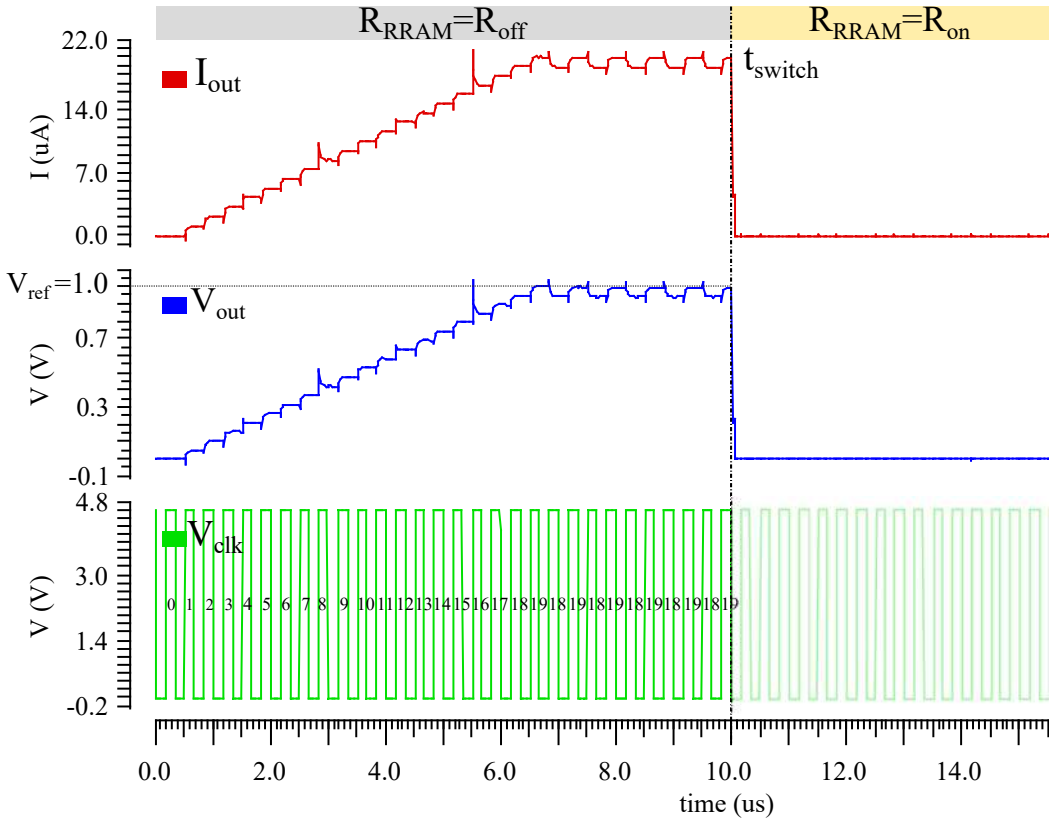


Figure 59 – Spice-simulation waveforms, showing the circuit operation as voltage regulator, so that output V_{out} equals input reference V_{ref}. When the set event takes place, $t_{switch} = 10\mu s$, the resistance drop is detected by the WT block and the system is shut-down.

4.3.1 *Current Switching Digital to Analog Converter*

A schematic view of the implemented DAC architecture is shown in Figure 60: the circuit consists in a tank of binarily-weighted current sources [129], controlled by the input code $S_{1:n}$ (for example the counter's output), which sum at the output node.
The activation/deactivation of each source, ranging from I_1 to I_n, is regulated by the closing/opening of corresponding switches $S1 - Sn$. The output current, I_{out}, results:

$$I_{out} = \sum_{i=1}^{n} 2^{i-1} Q_i I_1 \tag{31}$$

where n is the number of current sources, Q_i the logic value controlling the i^{th}'s switch (1=closed, 0=open), and I_1 is the smallest generated current. Figure 60 highlights the correspondence between each counter's output bit and current source: the weight is increasing from index 1 to n, 1 being associated to the counter's Least Significant Bit (LSB) and n to the Most Significant Bit (MSB). The circuit is

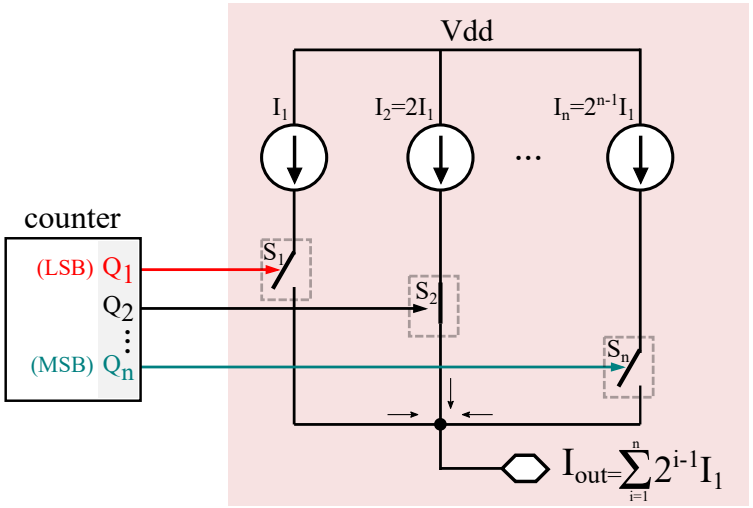


Figure 60 – Schematic view of the current-switching DAC circuit. The current sources $I_{1:m}$, are coupled to the output node if the corresponding switch logical value, $S_{1:n}$ is high.

designed so that the maximum amount that can be generated equals the RRAM technological current compliance, I_{cc}. From Expression (31):

$$\text{for } Q_i = 1, \forall i \in [1, n] \; : \qquad I_{cc} = \sum_{i=1}^{n} 2^{i-1} I_1 \tag{32}$$

As we targeted a I_{cc} in the range of 100µA, we achieved satisfactory precision for $I_1 = 1\mu A$, and $n = 7$. The chosen parameters, as better explained in the following Subsection 4.3.2, represent a compromise between regulation precision and circuit complexity, which trade which each other on the number (and topology) of current sources.

A TRANSISTOR-LEVEL IMPLEMENTATION is reported in Figure 61. Different amplitude are realized by mirroring the input reference, I_{ref}, at various W/L ratios (L being kept constant for every transistor). The i^{th} current source, I_i, can be sized given [129]:

$$I_i = \frac{W_i}{W_1} I_{ref} = m_i \cdot I_{ref} \tag{33}$$

where W_1 is the width of transistor M1 and $W_i = m_i W_1$ the width of the the i^{th} transistor [1]. In our implementation I_{ref} is externally supplied through a dedicated input pad; this choice allows to gain some flexibility in the generation of the output current by varying I_{ref}. In standard setting, $I_{ref} = 1\mu A$, and transistors M2 and M3 generate currents approximately equal to 80µA, namely $I_{ref\_Opamp\_OL}$ and $I_{ref\_Opamp\_driver}$, which are used to bias operational amplifiers Opamp_OL and Opamp_driver (not shown here). Transistors $M4 - M10$ stand for the actual

1. Equation (33) simplifies the real-case scenario, under the assumption of perfectly matched transistors ($L_1 = L_i$ and $V_{T1} = V_{Ti}$) and neglecting channel-length modulation.

current source tank, while M11 – M17 act as switches to either isolate or couple each source to the output node.

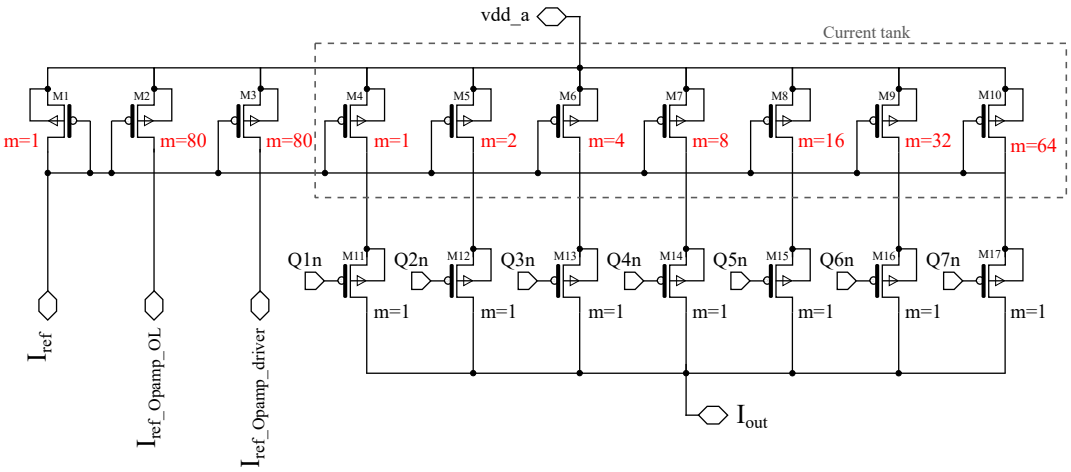


Figure 61 – Transistor-level design of the current generator, consisting in a switchable current mirror, where the LSB current, $I_{ref} = I_1$, is mirrored at increasing W/L ratios. For better matching, W is enlarged by putting multiple unitary transistors in parallel, the number being specified by the m factor

4.3.2 *Design Trade-Offs*

Figure 62 shows a time-transient Spice simulation of I_{out} and V_{out}. At the power-up, $I_{out} = 0$; it is then increased by $\Delta I = I_{LSB}$ (the inset shows a zoomed view) as the counter is incremented. Once the steady-state current, I_{ss}, is reached, the output remains afflicted by some ripple, as the count keeps oscillating. As the output voltage is proportional to I_{out}, V_{out} has a similar shape, and the current ripple translates into a voltage ripple around the regulation voltage, V_{ref}
The steady-state output voltage, V_{out}^{ss}, can be expressed as:

$$V_{out}^{ss} = (I_{ss} \pm \Delta I/2) \cdot R_L = V_{ref} \pm \Delta V/2 \tag{34}$$

where

$$\Delta V = \Delta I \cdot R_L \tag{35}$$

is the steady-state voltage ripple. Ideally, $V_{out} = V_{ref}$, so the output voltage is closer to the ideal case the smaller the ripple. Given Expression (35), this can be achieved either minimizing ΔI or R_L. Decreasing ΔI means decreasing I_{ref}, which in turn requires higher circuit complexity and area, as more transistors are required to obtain a sufficient amount of current to program a RRAM cell. Therefore, a trade-off between precision and complexity has to be made.
Moreover, the dependence of ΔV on R_L poses another design challenge, as RRAM variability brings high statistical load variation. In order to limit the spread, we

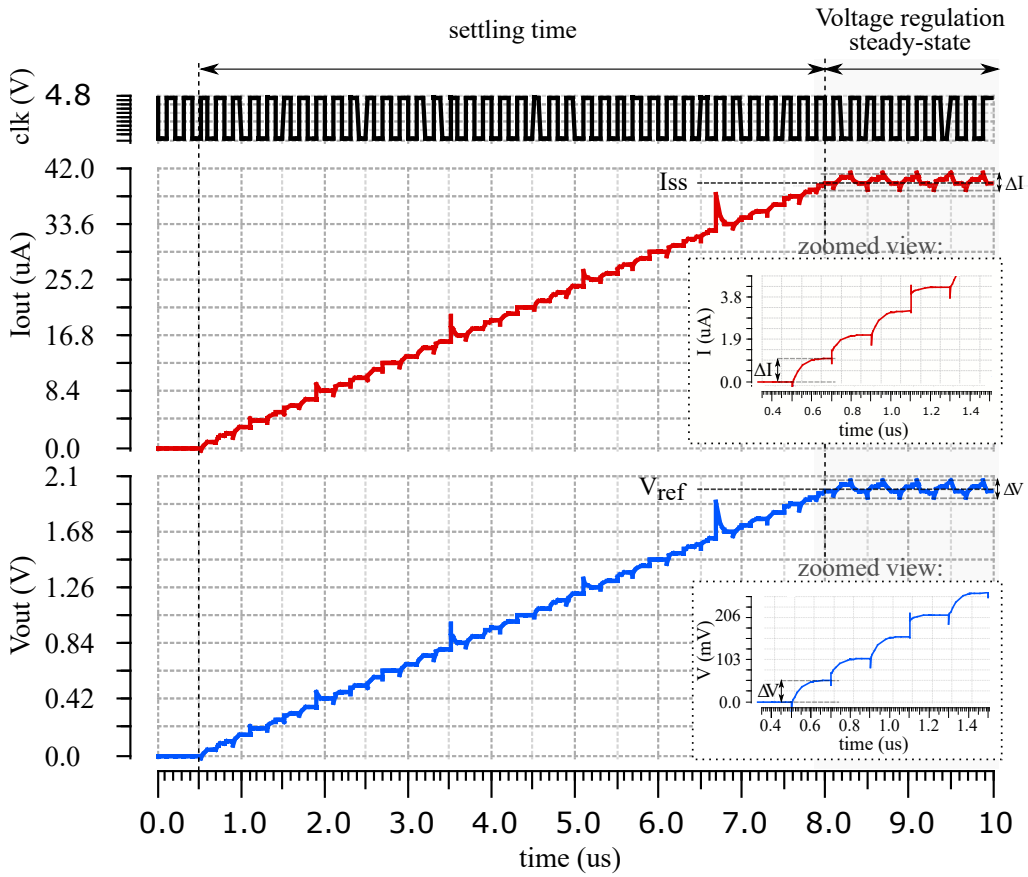


Figure 62 – Time transient Spice simulation of output current, I_{out} and voltage V_{out}. Values are incremented at each rising clock edge, until the device enters the Voltage regulation steady-state, where I_{out} and V_{out} settle around $I_{ss} \pm \Delta I/2$ and $V_{ref} \pm \Delta V/2$ respectively.

inserted a fixed parallel load, R_p, to the RRAM device. The steady-state voltage ripple thus becomes:

$$\Delta V = \Delta I \cdot (R_{HRS} \parallel R_p) \tag{36}$$

The worst case ΔV_{WC}, results when $R_{HRS} \gg R_p$, which maximises the load, and $R_L \cong R_p$:

$$\Delta V_{WC} = \Delta I \cdot (R_p) \tag{37}$$

Given Equation (37), the lower R_p, the lower the ripple. However, higher output current would be required to drop enough voltage, V_{ref}, over R_p. Moreover, the set process energy efficiency would be quite low, as a considerable share of I_{out} would be lost on R_p, instead of being used to switch the RRAM cell. Thus, sizing R_p represents another trade-off between regulation precision and energy efficiency. We posed $R_p = 100k\Omega$, which is a value that approaches the mean HRS state

for our technology. As a result, the worst-case voltage ripple becomes: $\Delta V_{WC} = 100mV$, and:

$$V_{out} = V_{ref} \pm 50mV \tag{38}$$

Considering that a typical set voltage is in the range of a few Volts, the steady-state output oscillates within a range of $\pm 5\% V_{ref}$, which is a satisfying voltage excursion for memory application.

4.3.3 *Digital Bidirectional Counter*

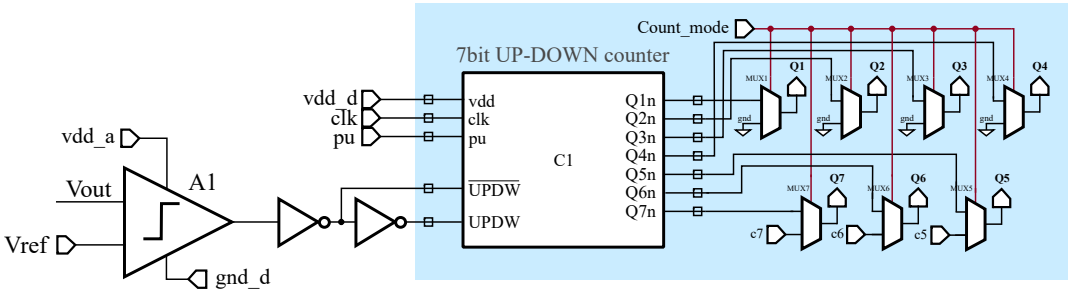
This subsection focuses on the Digital Counter's architecture and regulation circuitry. A schematic is shown in Figure 63a, where the central block represents the 7-bit bidirectional counter, C1. Figure 63b reports a lower-level view; seven JK Flip-Flops (FF) store each bit value, where the bit significance increases with the FF number: FF1 is associated to thr the LSB, while FF7 to the MSB.
In Figure 63c, a Spice time-transient simulation. At the power-up, $Q1 - Q7 = 0$; once trigger signal pu rises, the output starts increasing with each positive clock (clk) edge. Flag UPDW regulates the counting direction, which is positive when false, and negative when true. As V_{out} approaches V_{ref}, UPDW starts to oscillate, giving rise to the aforementioned steady-state ripple.
As shown in Figure 63a, the outputs of the whole counter block (Q1 : Q7) corresponds to the output of digital multiplexers MUX1 : 7. This solution allows to add an extra degree of freedom for both circuit operation and debug purpose: namely an *automatic* versus a *manual* DAC control. Count_mode allows to select either modes: C1 counter outputs (automatic operation), or external inputs (manual operation). In the latter configuration, the three most significant bits (c5 : c7) are coupled onto input pads, while the remaining (c1 : c4) are internally grounded.

4.4 SWITCHING DETECTION

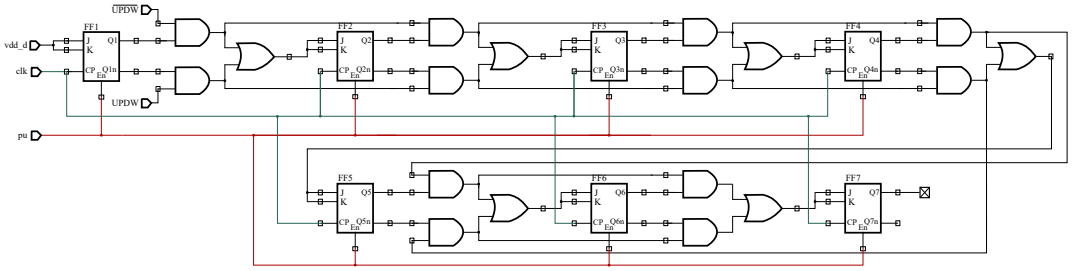
The ultimate task of the proposed circuit is to react to a set event, opportunely cutting the bias that damages the sample and gives rise to energy waste. The problematic has been already discussed in Chapter 3, where energy waste results from the high degree of variability of the memory switching times (see Section 3.1 for further details).
The following Sections present the self-terminating feature of the proposed circuit, as well as two energy-saving programming alternatives: a *Write Termination* and a *Write and Verify* protocol, where architecture complexity and set process speed trade with each other.

4.4.1 *Self-Terminating architecture*

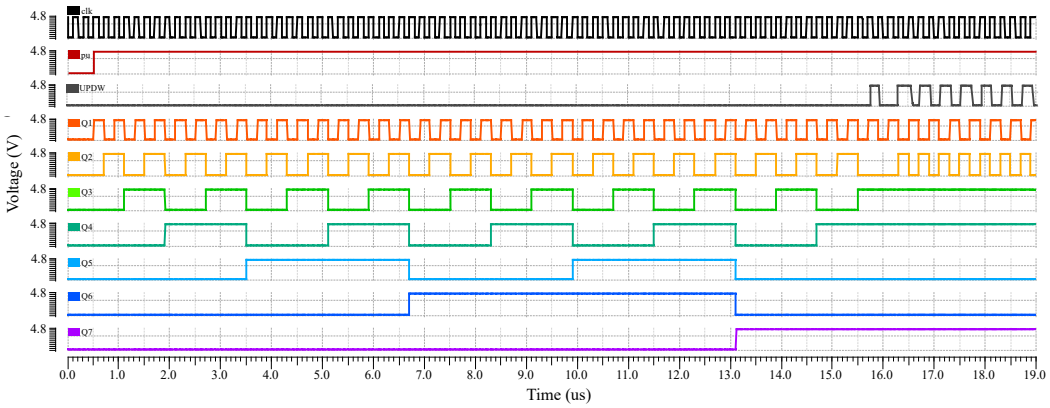
Introductory Section 4.1 presented our solution as intrinsically self-terminating, meaning that the circuit responds by itself to a sharp load drop, even in the lack of an external stop circuitry. Figure 64 illustrates this point with a qualitative analysis of how the circuit reacts to a set event.



(a) High level view of the bidirectional counter and annexed circuitry.



(b) Logic-gate level counter architecture: 7-bit up/down counting is realized by means of JK Flip-Flops: UPDW and $\overline{\text{UPDW}}$ inputs reverse the counting direction when toggled.



(c) Sice-simulation illustrating the control input and output states of the counter. Around 15.5µs, the counting direction keeps toggling as a result of entering the voltage regulation steady-state. As the output oscillates between 1001101 and 1001110, FF outputs Q1 and Q2 also continue toggling their state.

Figure 63 – Illustration of the bidirectional digital counter design and operation.

The voltage across the sample, $V_{out}(t)$, versus the time, is shown on top of the figure. The time scale starts at the settling time (t_{settle}), when the output voltage equals the reference V_{ref}; for clarity's sake, the steady-state voltage waveform is drawn ripple-less. In gray region (1) the cell is in HRS, while in yellow region (2), it is in LRS. The set event occurs, relatively instantaneously, at the interface between

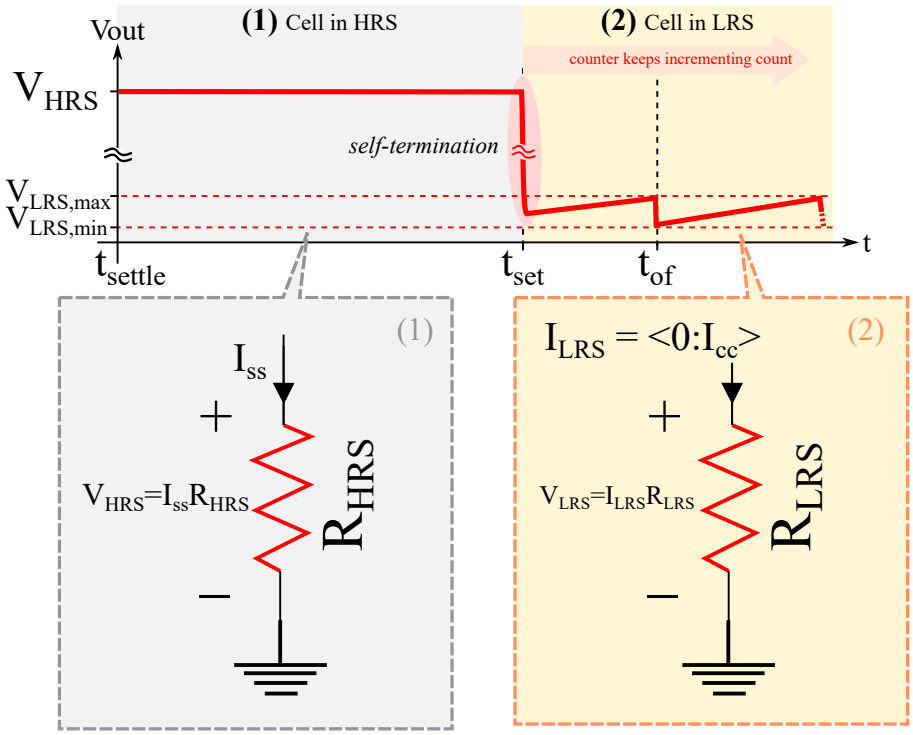


Figure 64 – Illustration of the circuit's *self-termination* mechanism, where a RRAM undergoes a set process that causes a fast load drop. In region (1) the memory is in HRS, while in region (2) in LSR; the output voltage V_{out}, represents the voltage across a memory sample.

the two region, at $t = t_{set}$. The voltage drop across the sample can be expressed as:

$$V_{out}(t) = \begin{cases} V_{ref} = I_{ss} R_{HRS} & \text{for } t < t_{set} \\ I_{ss} R_{LRS} & \text{for } t = t_{set} \\ I_{LRS} R_{LRS} & \text{for } t > t_{set} \end{cases} \tag{39}$$

where I_{ss} is the steady-state current, and I_{LRS} the current when the device is in set state, which ranges from 0 to the compliance value I_{cc}. Typically, $R_{LRS} \ll R_{HRS}$; therefore, as expressed by Equation (39), $V_{out}(t)$ dramatically falls as the resistance change takes place. As a result, the stress over the cell is reduced as soon as the programming operation is achieved.
Since the typical set voltage is in the range of a few Volts, $V_{out}(t_{set})$ falls down to hundreds of mV at most.
If the writing operation does not get interrupted, the counter control circuitry detects that V_{out} has again fallen below V_{ref}, and restarts incrementing the count. However, the maximum output voltage will be limited to $V_{LRS,max} = I_{cc} R_{LRS}$. Depending on the actual R_{LRS} and I_{cc} values, as well as the targeted V_{ref}, the regulation voltage might not be reachable after a set event. In Phase 2 of Figure 64, the typical case is shown, where R_{HRS} and R_{LRS} are several orders of magnitude apart, and V_{out} never reaches the regulation voltage once the memory transits to the set

state. As a result, the counter keeps counting forward, and its outputs eventually overflows at t_{of}, where the output is zeroed and V_{out} reaches a minimum.

4.4.2 *Write Termination Protocol*

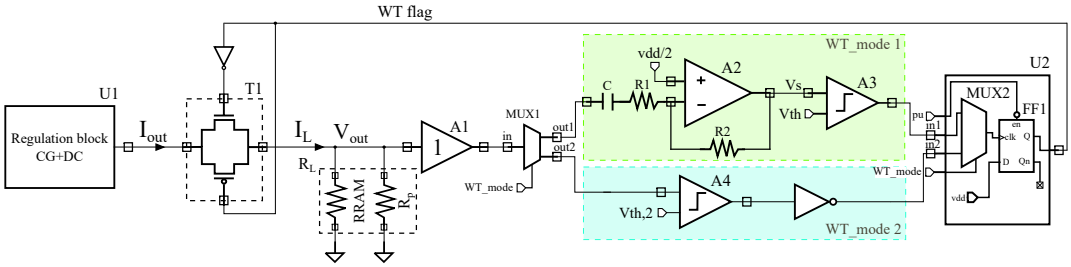
The in-built write termination feature does help reduce the stress over a device undergoing a set process, but cannot stop the considerable, unnecessary dissipation that still occurs once the transition is completed.
Therefore, a valid approach to save on energy expense is constituted by the Write Termination mechanism hereby presented.

WRITE TERMINATION CIRCUIT OPERATION Figure 65a illustrates the Write Termination mechanism present in our circuit. The load bias is interrupted by the opening of pass gate T1, which either couples or isolates the output of the DAC block, U1, to load R_L. If flag *WT*, corresponding to the output of block U_2, is low, T_1 is in closed position and $V_{out} > 0$, else $V_{out} = 0$.
Analog amplifier A1 buffers V_{out} in order to drive subsequent stages: namely, either out1 or out2 outputs of multiplexer MUX1, depending on its selection bit (*WT_mode*). *WT_mode* allows to choose between two different write termination circuitry: *WT_mode* 1 and *WT_mode* 2, which are highlighted in Figure 65a.
When $WT\_mode = 0$, *Termination Mode 1* is active; its operation is illustrated by the simulation waveforms reported in Figure 65b. In this topology, a set event is detected when a sharp output voltage transition takes place, and the high-pass filter constituted by capacitor C and resistor R_1 becomes conductive. If the amount of time required for a high to low resistance transition, δt_{set}, is $\leqslant 2\pi CR1$, out1 is coupled onto A_2's input [2]. Figure 65b shows that, at set time t_{set}, V_{out} sharply drops as a result of the load change. Correspondingly, signal V_s, amplified output of the high pass filter [3], becomes high enough to raise comparator A3's output. In turn, flag *WT* toggles its state to vdd, signaling a set event, and opening switch T_1.
Alternatively, by posing $WT\_mode = 1$, the switching can be detected through *Termination Mode 2* (*WT_mode* 2 block in Figure 65a). In this configuration, a set event is detected when the output voltage, which drives comparator A4's input, falls below threshold $Vth, 2$.

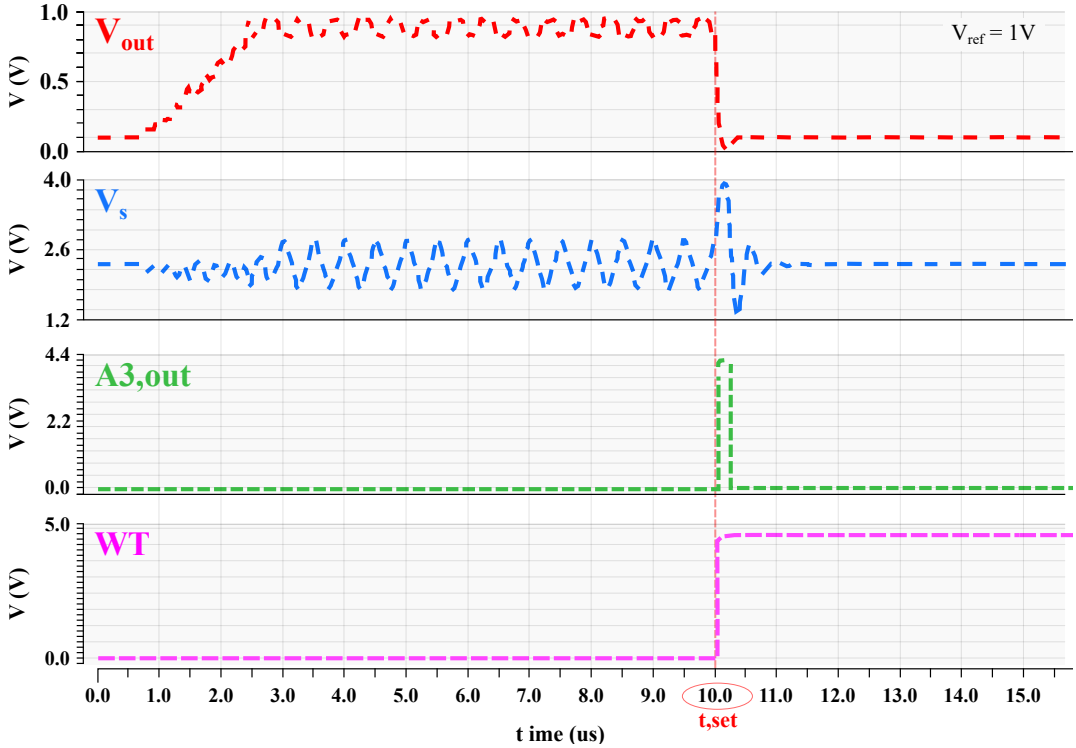
4.4.3 *Write and Verify*

A different approach to tackle the programming energy waste is constituted by the Write Verify (WV) method; an example illustrated in Figure 66. According to this solution, a set operation consists on a write pulse followed by a read pulse, which checks weather the previous attempt was successful: in such a case, the programming phase ends, else the cycle resumes [133–136].
By implementing short, amplitude-increasing pulses, it is possible to deliver discrete bursts of limited energy, hence avoiding considerable waste. The WV approach could also prove useful in lowering technological variability: for example,

2. As typical set switching times are in the range of $[1 - 100]$ns, C has been set equal to 5pF, while R1 to 20kΩ.
3. Resistor R_1 and R_2 set the gain of amplifier A_2: $G = -R2/R1$.



(a) Schematic of the write termination block, putting into evidence the two alternative strategies: *WT_mode 1* and *WT_mode 2*. The termination mode is chosen by configuring *WT_mode* digital input.



(b) Spice time-transient simulation of a set event, detected through *WT_mode 1*. At the set time, t_{set}, V_{out} sharply decreases, producing a large swing at node V_s (small oscillations are due to V_{out}'s steady-state ripple) which triggers a state change in *WT* flag, signalling a set operation completion.

Figure 65 – Write termination circuitry.

the sample's final resistance could be adjusted within some target range by modulating the amplitude/polarity of the programming pulses (as shown in Figure 66) [133].

The main inconvenience of this approach is the programming time inefficiency that results from the extended operational flow [133, 137]. Moreover, the resistance value so obtained is typically drifting over a short time interval, so that the reliability of Write and Verify methods has been questioned [136, 138]; constant refresh of the stored state might be required to ensure non-volatility, lowering the energy efficiency of this approach [133, 136, 137].

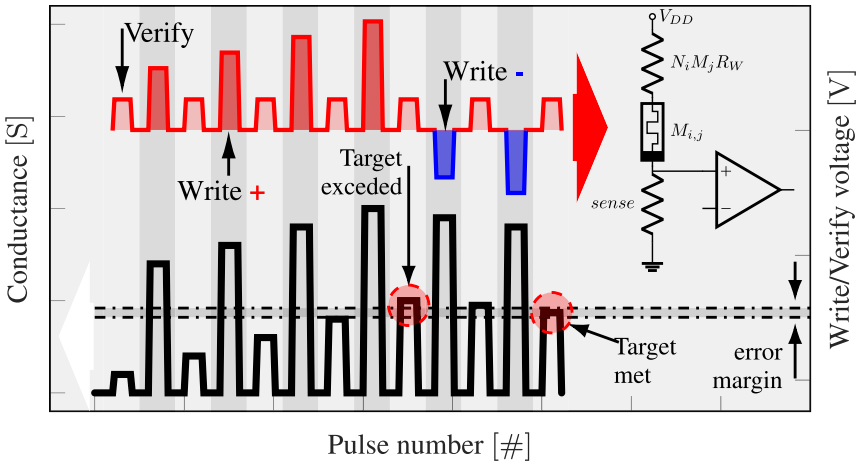
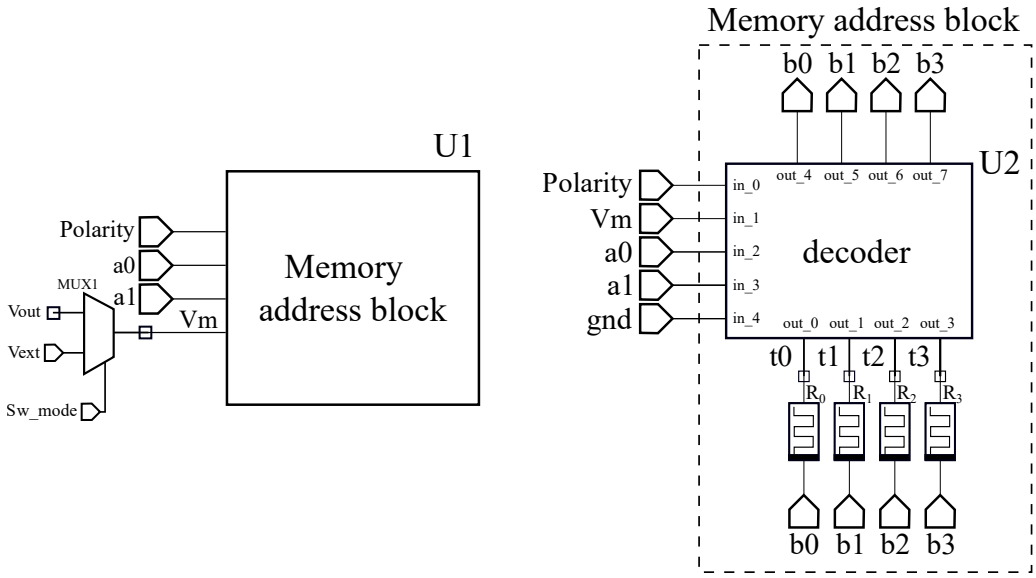


Figure 66 – Example of Write and Verify method, where each programming pulse is followed by a read operation. The amplitude of the write voltage is increased step-wisely, until the desired resistance value is achieved. Figure adapted from [133].

4.4.4 *Memory Addressing*

In the actual implementation, the load consists in a small array of 4 RRAM elements, which are selected through the circuitry shown in Figure 67a. Polarity, a_0, a_1 and Vext are externally-supplied inputs, while Vout is coupled to the output of the DAC block. Multiplexer MUX1 allows to select the voltage to be applied across the memory cell, between Vout and Vesxt. When Vext is chosen, the DAC block is bypassed, and the programming operation can be performed in the standard way (for example in in debug stage, or during a reset operation).
A lower-level schematic is shown in Figure 67b, where decoder U2 routes voltage Vm and gnd according to the configuration of the address bits (a_0, a_1, Polarity). The outputs of the decoder are the top ($t_0 : t_3$) and bottom ($b_0 : b_3$) electrodes of the four-cell memory array. Appendix Section 4.7.2 reports lookup tables illustrating the decoder operation in larger details.



(a) Top schematic view, illustrating the inputs of the memory address block.

(b) Lower level, view expanding the memory block. V_m and gnd signals are routed through decoder U_2 according to the address bit values.

Figure 67 – Memory address block and routing bias.

4.5 ELECTRICAL CHARACTERIZATION AND EXPERIMENTAL RESULTS

The circuit has been integrated on a 8-inch silicon wafer; a picture is shown in Figure 68. A microscope photograph, as well as an overlay of the full circuit layout, highlighting the main constitutional blocks, can be seen in in Figure 69. The front-end was designed and fabricated in 130nm bulk CMOS technology, operating at 4.6V nominal supply voltage. The bottom half of Figure 69 shows a TEM picture highlighting the memory integration as back-end process, where RRAM memory cells are sandwiched between M4 and M5 metal levels. The memory stack is TiN/Ti/HfO_2/TiN, where the HfO_2 oxide thickness is 5nm. The memory cell size is 300x450nm, and the footprint results 0.135μm^2.
The following Sections present the results of experimental characterization of the main circuital blocks, namely: the output buffer 4.5.1, the parasitic access impedance brought by the multiplexers 4.5.2, and the programmable current source 4.5.3. Lastly, Section 4.5.4 shows and comments tests on RRAM samples, where the impact of the designed architecture on the memory technology is evaluated.

4.5.1 *Characterization of the Output Analog Buffer*

In order to drive the capacitive load of the circuit's output pad, V_{out}, a unity-gain amplifier is put in place (see instance A_1 in the schematic shown in Figure 65a). To ensure that no additional signal distortion arises from the buffer itself, its static response needs to be characterized. Figure 70 shows the output over the Full Voltage Range (FVR), for two different samples (each situated on a different wafer die). As the circuit is biased between Vdd and gnd, the output is not rail-to-

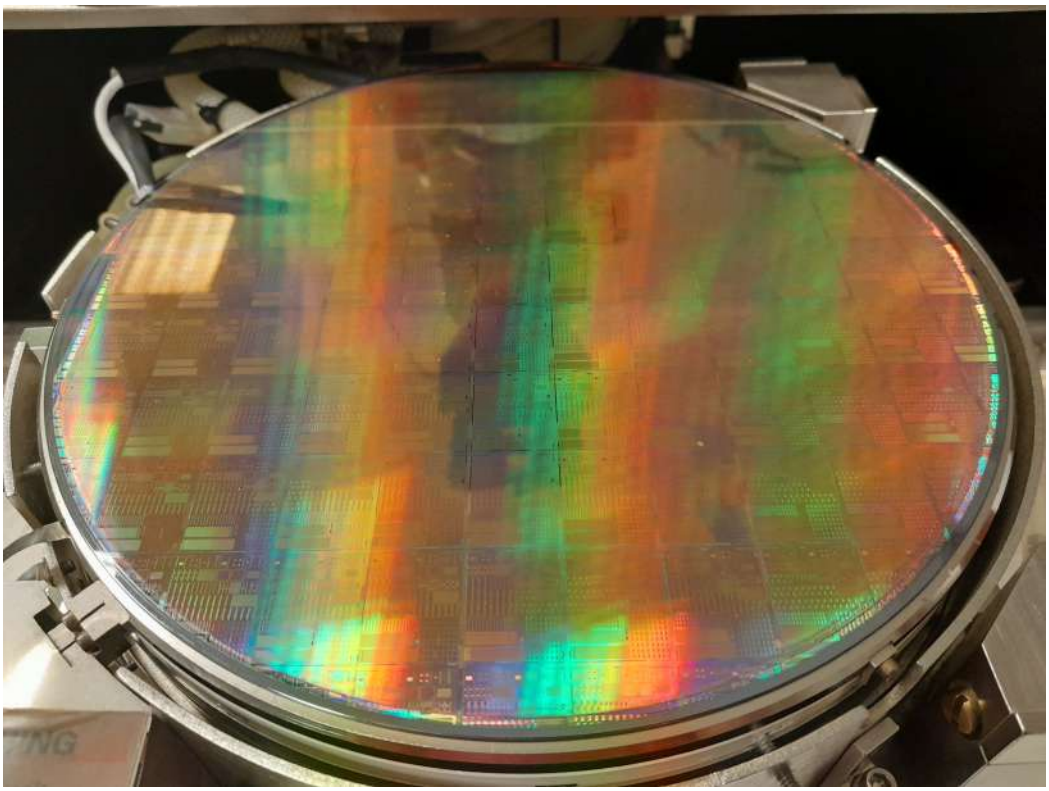


Figure 68 – Silicon wafer, onto which the designed circuit was integrated.

rail, with an upper saturation level that is found to vary depending on the DUT. Nevertheless, the block looks functional, and the distortion level appears negligible within the expected operational range, as typical set voltages fall between $1-3V$.

4.5.2 *Evaluation of Access Impedances*

In order to assess the reliability of read and write operations, we evaluated the parasitic contributions affecting the load impedance. Figure 71a shows a schematic with the overall resistive and capacitive elements that lie on the path of a memory cell. Their values need to be quantified in order to calibrate the experimental raw data, and deliver precise memory characterisation. The memory cell, highlighted in Figure 71a, is non-standard: it consists of a memory cells with two series transmission gates, one at each electrode. This design choice was made in order to reduce the parasitic series resistance; as our circuit operates as a controlled current sourced, there in no need to put any additional limiting elements.
Series resistor $R_{on,MUX}$ is the on-impedance of the multiplexer, causing a reduction of the voltage that ends up dropping on the RRAM. R_{IC} is an integrated poly-silicon resistor, by design equal to $100k\Omega$, in parallel to the memory cell to protect it from over voltages. On the downside, as the RRAM values approaches that of R_{IC}, the energy dissipated by the memory, as well as its read resistance, become affected. Parasitic capacitor C_{dev} influences the dynamic response of the circuit and contributes to the overall energy delivered to the memory, as is further discussed in Section 4.7.3. Figure 71b shows the evaluation of the multiplexer on

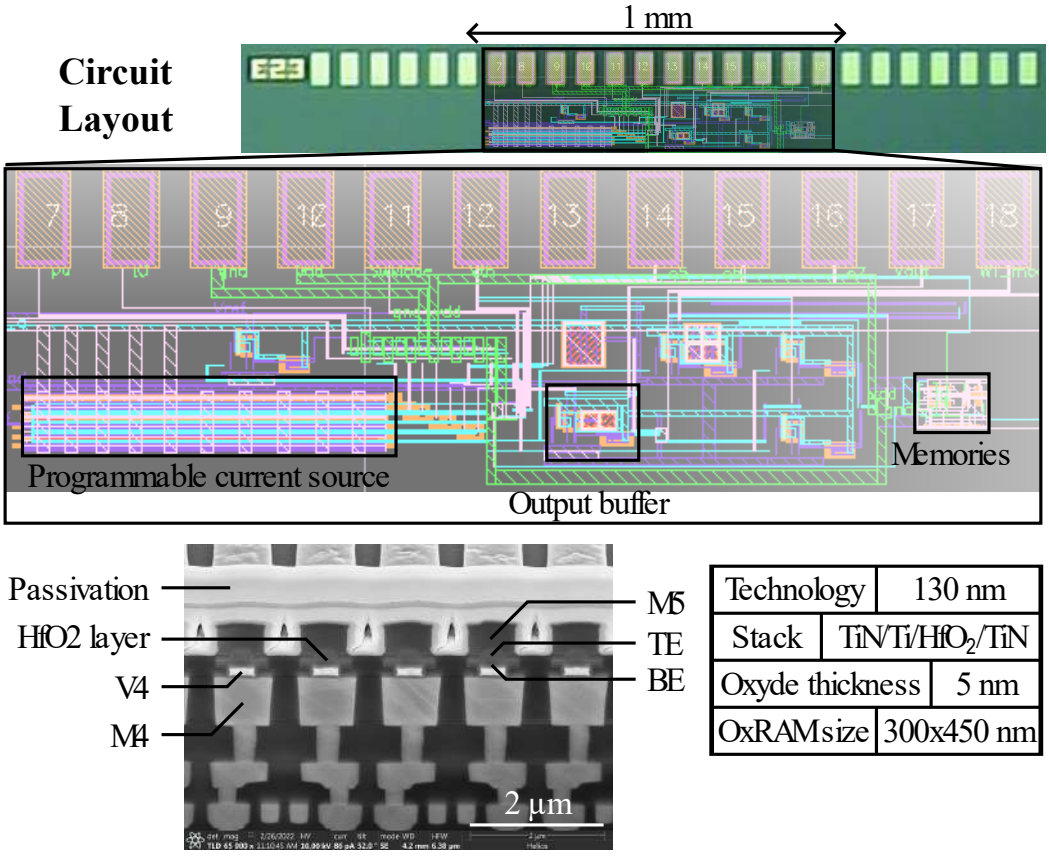


Figure 69 – On top, the full circuit layout, highlighting the main design blocks. The bottom-half of the figure shows a TEM picture where the back-end integration of the memory cells is in evidence.

resistance, versus the common-mode input voltage. The inset of the figure shows the test setup; the multiplexer is not connected to the DAC current source I_{DAC}, and TE pad is grounded in order to avoid loading by R_{DEV} and R_{IC}. Capacitance C_{dev} was extracted by post-layout simulation, and found equal to 600fF. At last, once $R_{on,MUX}$ is known, R_{IC} was derived by applying a constant input voltage through an SMU unit, V_{in}, and monitoring both the current drawn and V_{out}, in a layout topology where the memory element is missing ($R_{DEV} = \infty$). At $V_{in} = 1V$, $R_{on,MUX}$ was found equal to 1.89kΩ and $R_{IC} = 97.5k\Omega$. Once their values were quantified, a memory cell resistance could be read coupling the DAC current source and monitoring output pad V_{out}: Figure 71c shows the evaluated value versus the DAC digital input.

4.5.3 *Characterization of the I-DAC*

The core block of the designed architecture consists on the current DAC, which supplies the bias for the memory programming operation. Therefore, the primary objective of our tests was to evaluate the performance of the programmable current source. To this end, we analyzed the static errors that deviate the actual characteristic from the ideal case, namely: the offset and gain errors, as well as the integral and differential non-linearities. It is noteworthy to mention that the designed

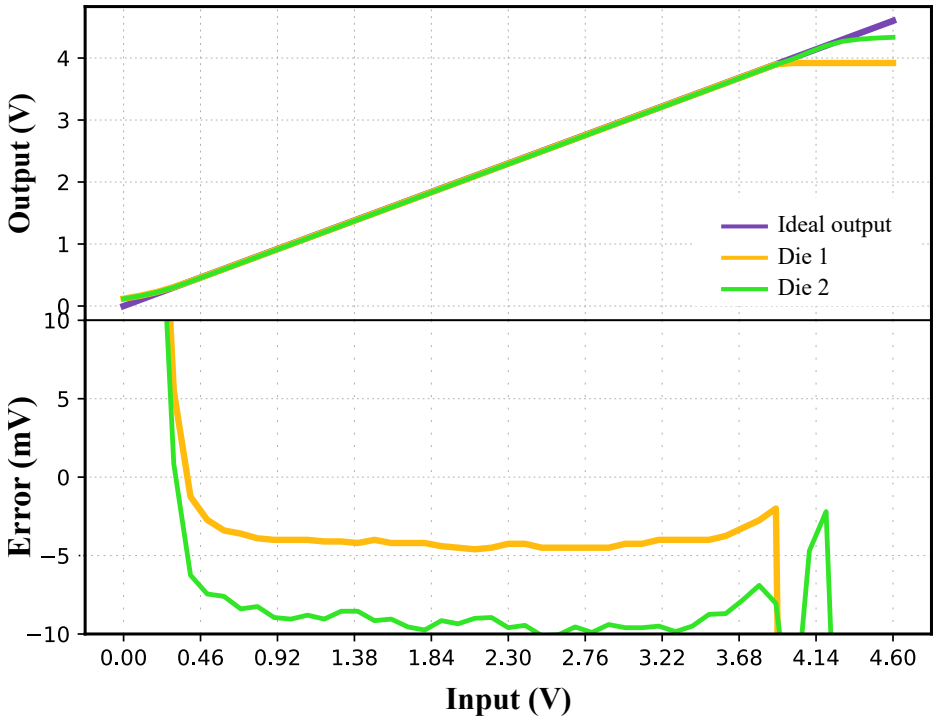
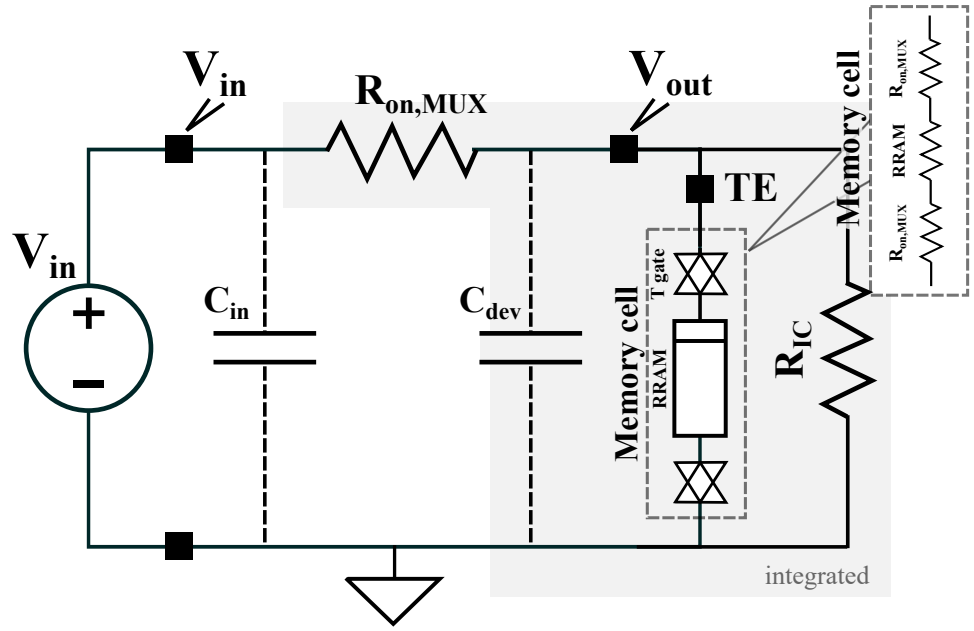


Figure 70 – Characteristic of the buffer driving the circuit's analog output pad, V_{out}, which follows the DAC's output voltage. Two sample characteristics are shown here, versus the ideal case. As the input is varies across the FVR, the amplifier outputs track the source, until saturation bends the curve at the power rails.

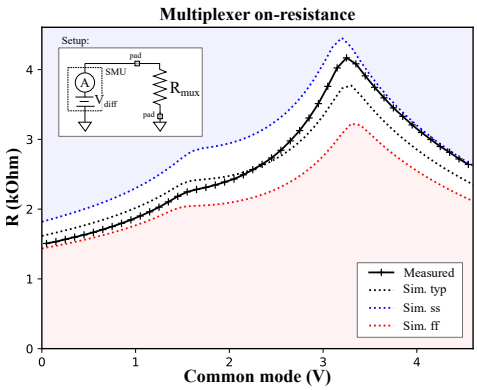
DAC architecture, which is implemented in the original context of (digital) memory programming, does not demand strict linearity specifications. Therefore, our objective is to assess the functionality and transfer characteristic of the designed circuit, rather than bench-marking our DAC with SoA typologies.

OFFSET AND GAIN ERRORS. Figure 72a illustrates the *offset error*, which corresponds to the resulting analog output when the input code is zero [132, 139–141]. An offset error appears when the output value is non-zero, causing a vertical displacement on the transfer function.
In Figure 72b, the *gain error* is shown. It is defined as the difference between the full scale ideal and actual analog output (Full Scale Range, FSR), obtained when the input code is at its maximum value (Full Scale Code, FSC) [132, 139–141]. Complementary details on the offset and gain errors are reported in Appendix Figure 4.7.2. Since all input codes are equally affected by offset and gain errors, it is possible to perform a calibration of the raw characteristic to mask their effect: Appendix Figure 4.7.2 shows a typical procedure for error-correction, which we adopted for our samples.

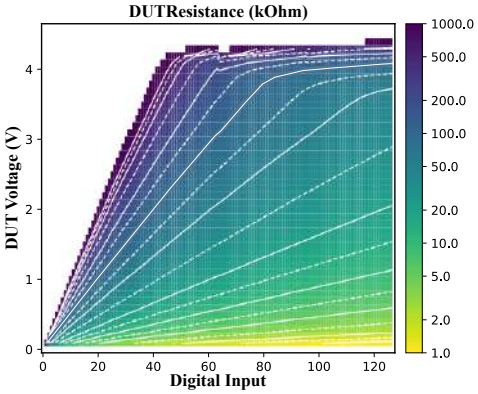
NON-LINEARITY ERRORS. Non-linearity errors cause a drift in the linearity of the DAC transfer function which, ideally, is a straight line. Two main types are defined, the *Differential Non Linearity (DNL)* and the *Integral Non Linearity (INL)*. A DNL occurs when the step between the outputs of two adjacent input codes is greater, or inferior, than a LSB (the ideal cause being exactly 1LSB) [132, 139–141].



(a) Schematic of test setup for parasitic contribution characterization. The memory cell, highlighted, consists of a RRAM in series with two transmission gates, located at the top and bottom electrode, whose resistance equals that of the analog multiplexer. The integrated circuit part is shown over a gray background; the black squares represent the access pads.



(b) Multiplexer's on-resistance versus the input voltage common mode. Monte Carlo simulations, as well as actual measurements are shown. The inset reports the test setup.

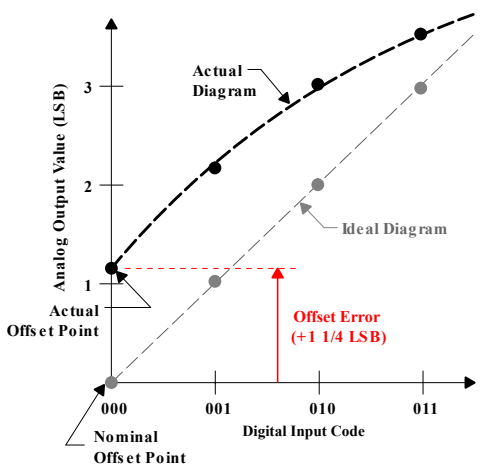


(c) A DUT resistance can be indirectly measured by monitoring its voltage, at a particular DAC input code.

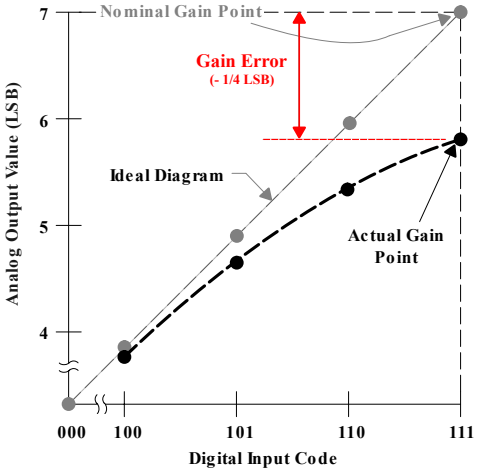
Figure 71 – Experimental evaluation of parasitic contributions.

Figure 72c shows an example of DAC characteristic affected by DNL errors. If the DNL < -1, the transfer function can become non-monotonic[4] [132, 139], while if DNL > 1, missing codes might occur [132, 139].
The INL, illustrated in Figure 72d, is a similar concept to the DNL, but it is defined as the maximum distance from the actual transfer function to the ideal straight line.

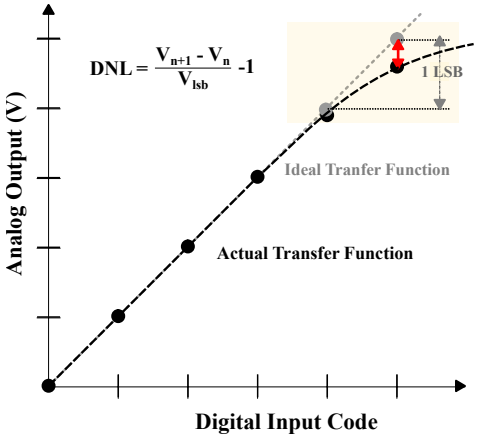
4. a non-monotonic curve results when the magnitude of the output becomes smaller, at an increase in the magnitude of the input[132].



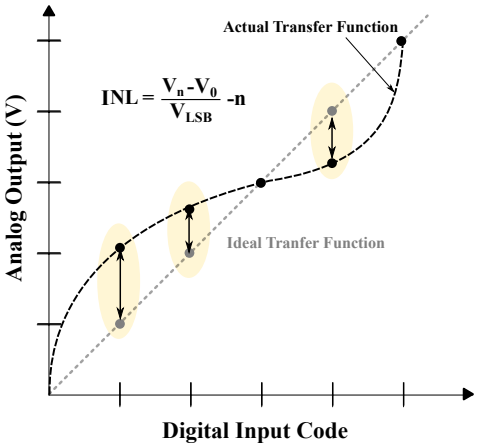
(a) Offset error, expressed as the difference of analog outputs between the ideal and actual characteristic, when the input code is zero. Adapted from [132].



(b) Gain error, which is evaluated between the ideal and actual characteristics at full-scale input code. Adapted from [132].



(c) Transfer function affected by negative DNL, which causes the reduction of the step output voltage. Adapted from [139].



(d) INL error, highlighted in figure, which measure the deviation from the linear transfer function. Adapted from [132].

Figure 72 – Static errors in DACs.

Non-linearities errors are evalued after the DAC characteristic is gain and offset error corrected.

EXPERIMENTAL EVALUATION OF STATIC ERRORS Figure 73 reports the DAC transfer function (analog output current I_{out} versus the input code), as well as the measured static errors afflicting the circuit. More samples are characterized in Appendix Section 4.7.5. Specifically, the offset error results negligible, while the gain error amounts to 4 LSB, or 0.03 FSR (FRS = 127µA). Both DNL and INL errors remain strictly below 1 LSB, reaching a max of 0.3 and 0.26 LSB respectively, ensuring good linearity of the transfer function and strict monotonic behaviour. Table 8 reports a summary of the main specifications and performances. From our analysis, we can conclude that the DAC is functional and behaves accordingly to

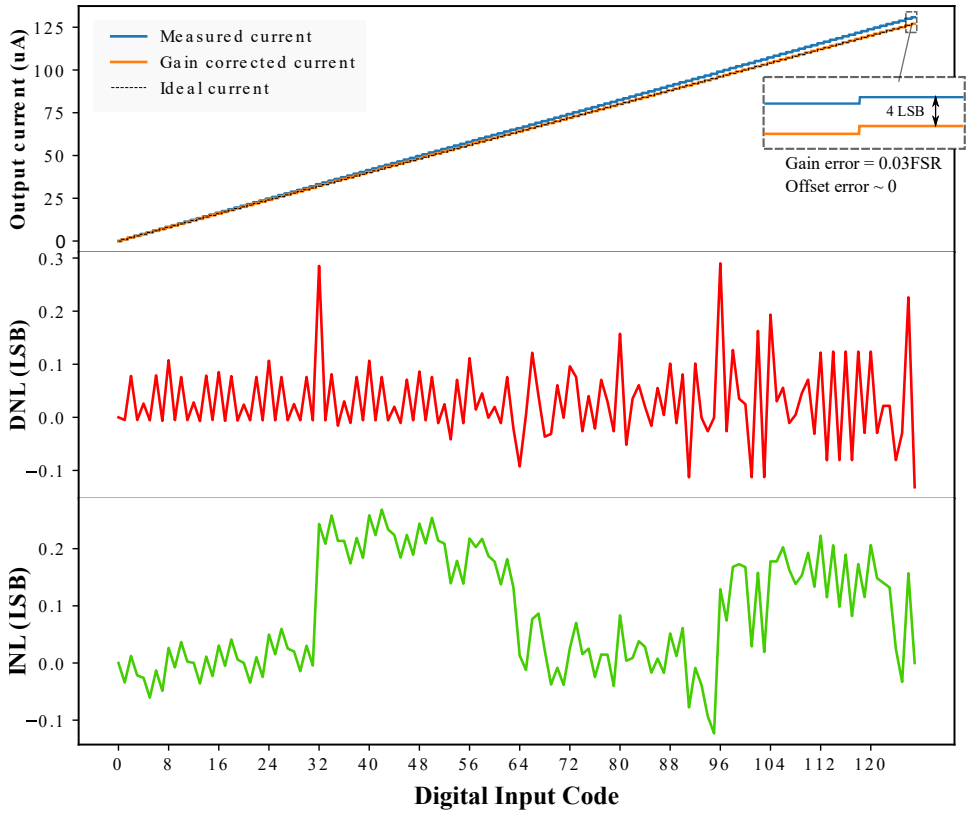


Figure 73 – Static error characterization for the integrated DAC block, where the INL is calculated on the gain-corrected characteristic. Overall, the DAC shows good static performances.

expectations. Moreover, the evaluated static performances are satisfying and in line with general purpose architectures.

4.5.4 *Current-based set process*

As explained in the introductory circuit overview, the designed architecture offers the possibility of being operated in two different modalities: as a current or as a voltage source. The former can be obtained in open-loop configuration, where the DAC digital input controls the output current. Figure 74 shows an equivalent circuit during a set operation, where the DAC output is the programmed current, I_{set}. Qualitative time transients of the current and voltage across the memory are also reported. In the absence of a Write Termination mechanism, current I_{set} continues to bias the RRAM until the end of the programming pulse. However, when the memory resistance drops, it voltage also does: the system self-termination mechanism lowers the energy waste and cell damage. Moreover, since a high impedance, current-limiting transistor is missing in our design, the programming operations are more energy efficient than standard 1T1R cells, at an equal programming time, thanks to reduced, unwanted voltage drop over the series impedance.
Therefore, we begin our experimental evaluation on the premise that a current driven set process looks appealing when considering the low energy waste even

Parameter	Designed	Tested			Unit
		min	typ	max	
Bit number	7	-	-	-	Bit
LSB	1	0.62	1.1	1.31	μA
FSR	127	130.7	130.9	131.8	μA
Resolution	0.008	0.01	0.009	0.008	FSR
Output V	>3.5	3.92	4	4.35	V
Offset Error	0	0	0.01	0.3	LSB
Gain Error	0	3.7	3.9	4.8	LSB
DNL	0	0	0.11	−0.62	LSB
INL	0	0.01	0.15	0.40	LSB

Table 8 – Main specifications, designed and tested, of the implemented IDAC.

in the absence of a WT mechanism, possibly improved memory performances and longer lasting cells.

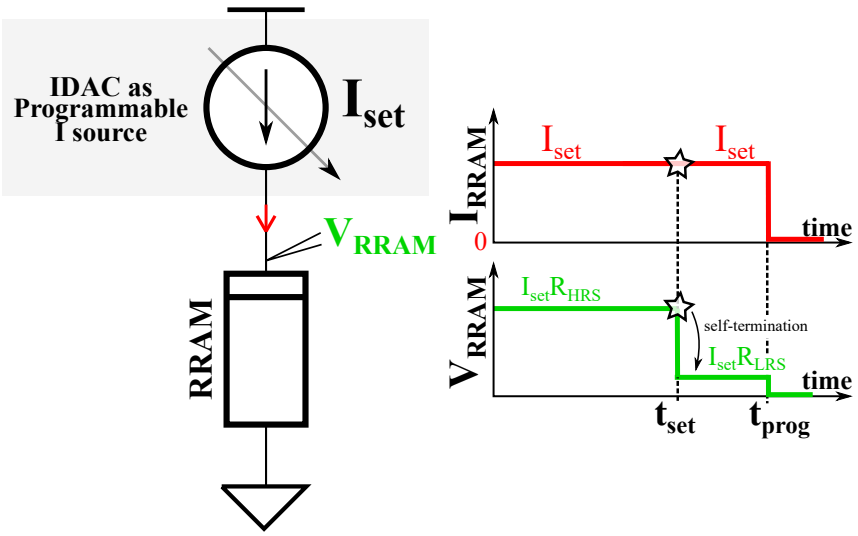
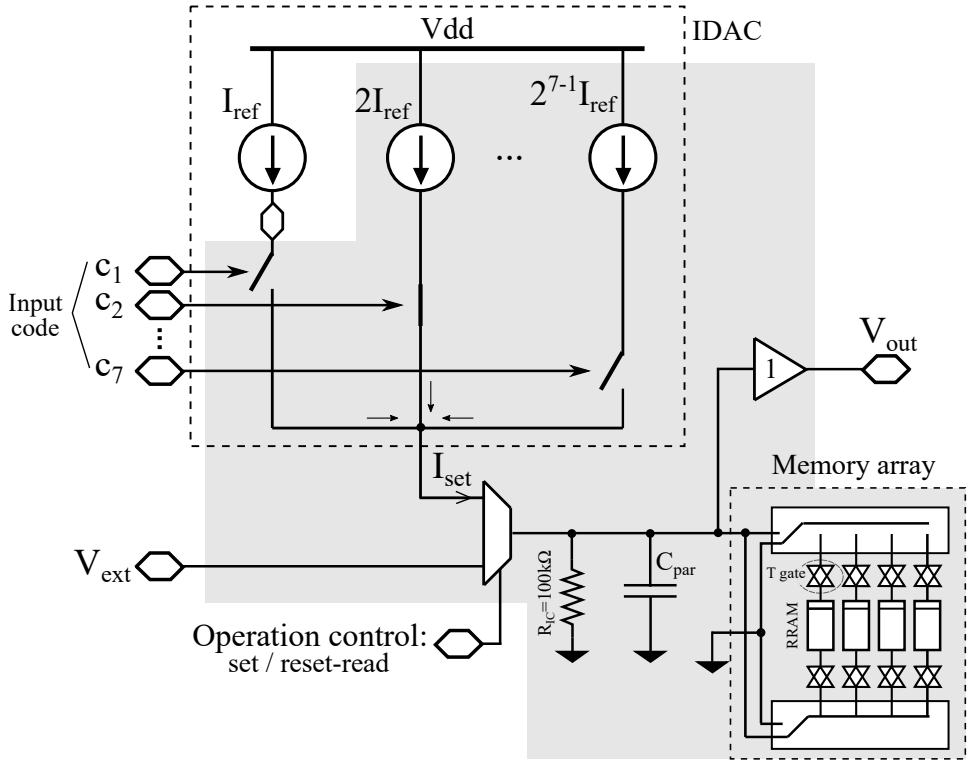


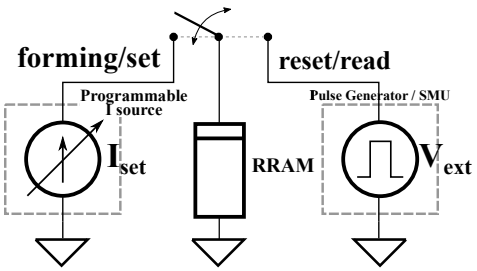
Figure 74 – Equivalent circuit illustrating the operation as programmable current source. The current sourced by the IDAC, I_{set}, remain constant after the set process occurs, so that the voltage falls as the resistance drops.

EXPERIMENTAL CHARACTERIZATION OF MEMORY SAMPLES The tested circuit is illustrated in Figure 75a. The forming and set processes are performed through current pulse I_{set}, supplied by the IDAC block, while the reset and read operations are performed by external voltage source V_{ext}. According to the value of digital input $c_1 : c_7$, the amplitude of the set current can be varied. The memory array, comprising of 4 cells (per die), allows access to one cell at a time. Depending on the address bit configuration, the routing and polarity can be varied (a detailed explanation is given in 4.4.4). Transmission gates at both the top and bottom elec-

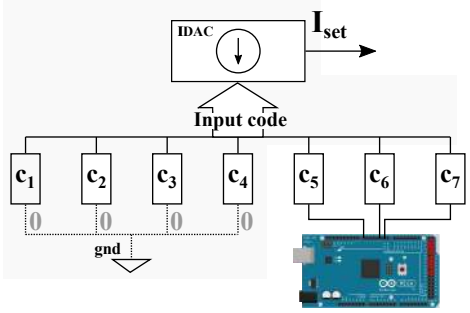
trodes allow good signal integrity and low voltage drop over the access elements. Figure 75b shows a simplified view of the different programming processes; in our setup, the reset voltage was 2V. while the read 0.2V. The circuit is operated in *Manual Mode* (as described in Section 4.3.3), i.e. not driven by the counter block. Figure 75c shows the setup for the control of the IDAC input bits, where the 3 Most Significant Bits (MSB) of the digital code are supplied by an Arduino microcontroller.



(a) Schematic showing the circuit operation as programmable current source. The gray area highlights the integrated part. During the set/forming operation, the current delivered by the *programmable current driver*, I_{set}, is coupled across the selected memory cell. For the reset and read operations, external voltage V_{ext} is selected instead.

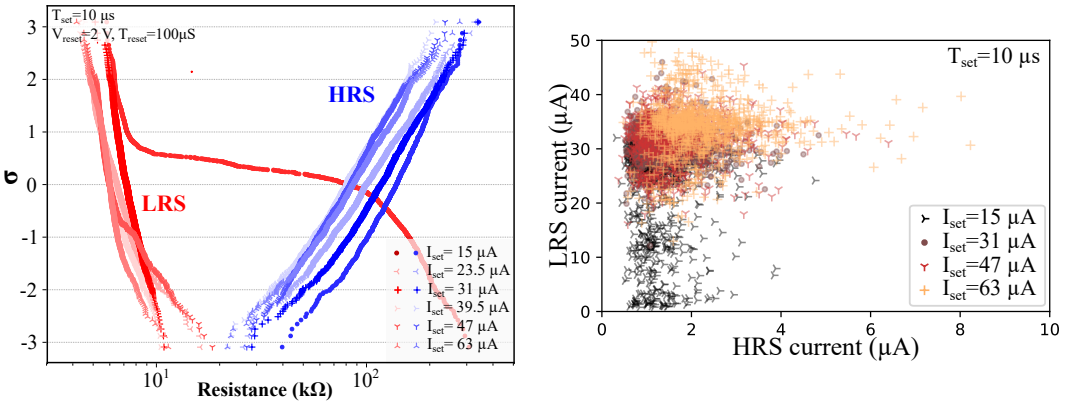


(b) Simplified diagram, illustrating the operation of the circuit during set/forming and reset/read processes. V_{ext} is externally supplied by a Pulse Generator (reset) or a SMU unit (read).



(c) Setup for current control of the I_{prog}, according to the circuit operation in *Manual Mode*. The 3 MSB are supplied by an Arduino Micro controller, while the remaining are internally grounded.

Figure 75 – Test setup for current-driven set/forming process.



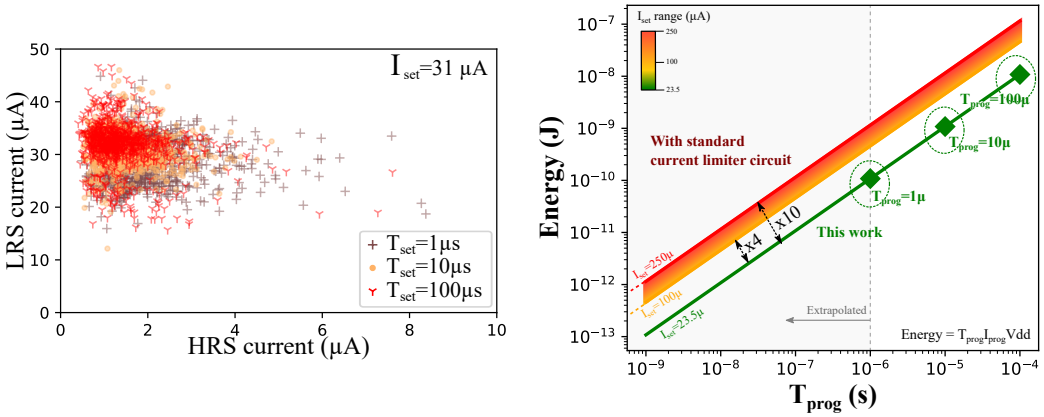
(a) Set process at different current amplitudes. LRS (in red) and HRS (in blue) resistance distributions are shown for a population of 500 samples. Above a threshold of 23.5µA, overlapping satisfactory results are obtained.

(b) HRS versus LRS current, for different programming set current amplitudes. Above 15µA, limited improvement can be achieved.

Figure 76 – Effect of the programming current amplitude on the set process.

IMPACT OF THE SET CURRENT AMPLITUDE Figure 4.5.4 shows the results collected on a current-driven set operation, while varying the programming current amplitude. I_{set} was ranged between 15µA and 63µA, adopting the setup described in the previous Paragraph 4.5.4. The programming pulse duration, for set and reset, was fixed at 10µs and 100µs respectively. An external voltage of 2 was used to perform the reset operation, while a reading pulse of 0.2 V was supplied for the read. In Figure 76a are reported the resulting LRS and HRS resistance distributions. while in Figure 76b the LRS and HRS currents are shown. It can be seen that, for a programming amplitude of 15µA, the distributions are heavily overlapping. On the other hand, no further real performance gain is obtained above 23.5µA, as the resulting distributions are overlap one another. We conclude that the WM cannot be increased at the expense of higher programming energy, and thus the optimal choice is $I_{set} = 23.5\mu A$, which allows to minimize the process consumption. Further results, which confirm this trend are reported in Appendix Section 4.7.6 and Section 4.7.6.

IMPACT OF THE SET TIME DURATION Figure 77a shows LRS and HRS distributions, obtained by ranging the duration of the set current pulse while keeping the amplitude constant to 31µA. It can be seen that there is no appreciable difference between different trials, whose programming time was varied between 1µs and 100µs, the former being the shortest time the setup allowed to reach. Combining the results obtained on both the impact of amplitude variation and programming duration, the shortest set pulse at the lowest acceptable amplitude ($I_{set} = 23.5\mu A$ and $T_{prog} = 1\mu s$) is the optimal choice in terms of power consumption. Figure 77b shows the set energy consumption versus the programming time, estimated at the first order as $T_{prog} I_{set} Vdd$ [116]. Our work (curve in green) is compared to SOA architectures (shaded orange-red area), where the current compliance is typically in a range between $100\mu - 250\mu A$ [68, 142–144]. It can be seen that the proposed circuit consumes less energy, and the consumption gain sits between a factor of



(a) Effect of different current pulse duration. No variation of the HRS and LRS currents can be appreciated upon varying the programming between 1µs and 100µs.

(b) Energy consumption per set operation. Comparison between this work (in green) and the SOA (orange to red shade) where the current compliance is varied between 100µA to 250µA.

Figure 77 – Effect of programming pulse duration.

4 to 10. As our experimental tests demonstrated that the final resistance distributions is uncorrelated to the pulse duration (Figure 77a), set times ($< 1\mu s$) are envisage-able. Therefore, an even lower energy cost can be targeted (extrapolated area in Figure 77a).

IMPACT OF THE PARASITIC LINE CAPACITANCE Figure 75a shows capacitance C_{par}, in parallel to the integrated 100kΩ resistor. As previously discussed in Section 4.5.2, this element is not present by design, and appears due to the parasitic capacitance of the metal lines. Figure 78 gives a graphic illustration of the effect of capacitor C_{par} during a set process.
The parasitic capacitance appears in parallel to the memory cell. Therefore, when the voltage tends to drop at set time (t_{set}), C_{par} holds this value until discharged onto the RRAM. As a result, a current overshoot flows across the sample; its peak amplitude is proportional to the conductance jump R_{HRS}/R_{LRS}, and it is thus significant (generally in the mA range). However, its time duration is proportional to $C_{par}R_{LRS}$, and hence too short (typical τ is in the low ns range) to cause damage to the cell.
A shown in Figure, C_{par} contributes to the energy that is overall delivered to the memory. The increase is half the energy stored inside the capacitor, while percentage of energy contribution by C_{par} to the overall amount, ΔE, results approximately equal to:

$$\Delta E = \left[1 + \frac{2t_{set}}{C_{par}R_{HRS}}\right]^{-1} \tag{40}$$

where E_C is the energy delivered by C_{par}, E_{ideal} the energy consumed when $C_{par} = 0$, and t_{set} the set time (details of calculation are reported in Appendix Section 4.7.4). For t_{set} in the range of $10ns - 1\mu s$, ΔE results, nominally [5], between

5. Estimation derived using $C_{par} = 600fF$ and $R_{HRS} = 100k$.

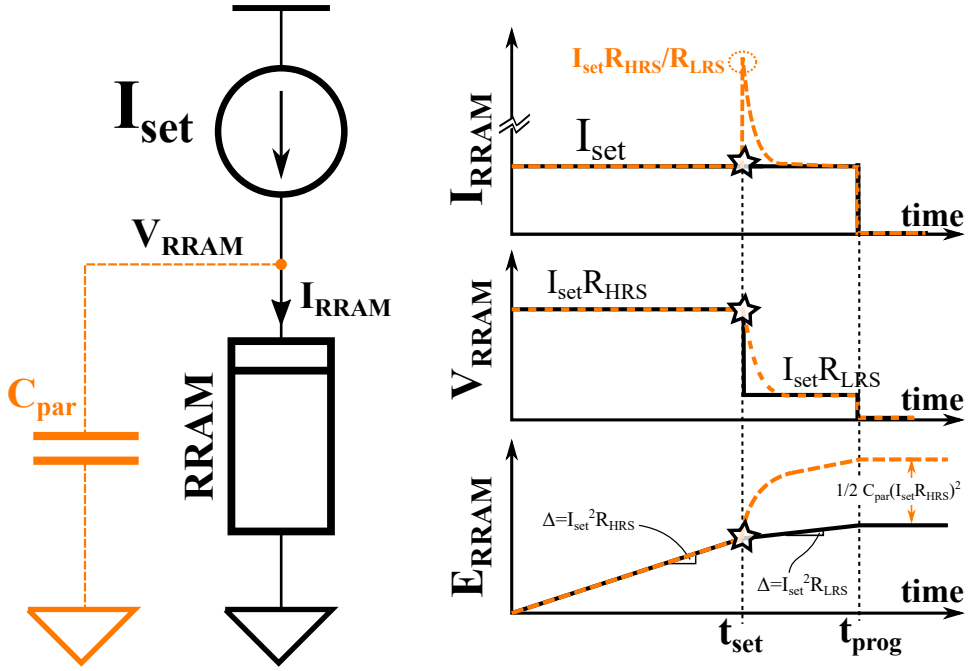
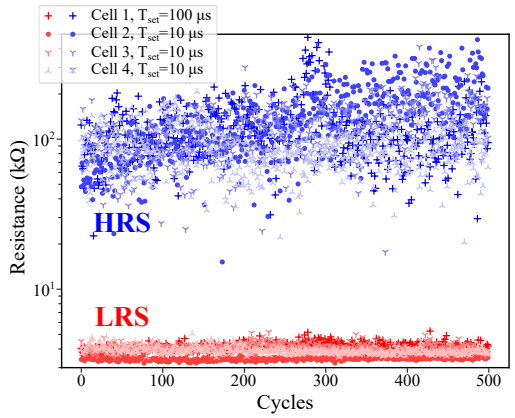


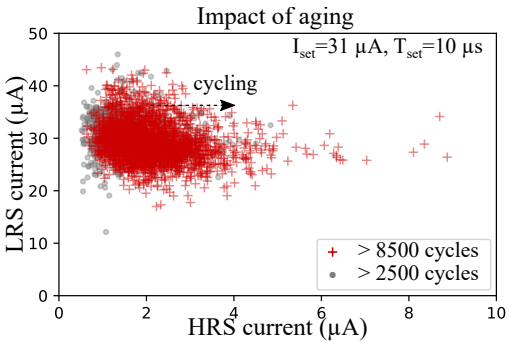
Figure 78 – Effect of parasitic capacitor C_{par} during a set process. In black, transients for the case when C_{par} is negligible, while in orange the effect of C_{par} on the current, voltage and energy dissipated over the RRAM cell.

3% to 75%. As a result, C_{par} can significantly affect the switching process, and it becomes the dominant contribution for the shortest RRAM set times. This conclusion should not sound surprising, after the discussion presented in Chapter 3 on the idea of programming RRAM by a charged capacitor.
After clarifying the role of C_{par} in our circuit, we could give interpretation to the results collected so far, considering its influence. The low sensitivity of set performances upon the variation of the programming current can be explained by the fact that the energy delivered by the capacitor makes the RRAM less sensitive to a change of I_{set} alone. Nonetheless, a *threshold* exists (for $I_{set} < 23.5\mu A$), below which the capacitor cannot deliver enough energy to achieve a satisfying LRS distribution.
A similar explanation can be suggested regarding the independence upon the programming time, where C_{par} is seen as the main contributor.

ENDURANCE ANALYSIS Figure 79 shows the results of endurance tests, carried out with a set current of 31µA, at various programming times, and reset voltage of 2V for 10µs. In Figure , 500 set and reset cycles are shown: the mean WM is equal to 25, and remains rather constant until the end. In Figure 79b the impact of aging is inspected; HRS versus LRS currents are shown for one sample, over 11k cycles. In gray are reported 1000 cycles after the first 2500, while in red other 2724 after 8500 cycles. It can be seen that the distribution tends to move to the right when increasing the cycle number, as the HRS gradually collapses onto the LRS, closing the WM.

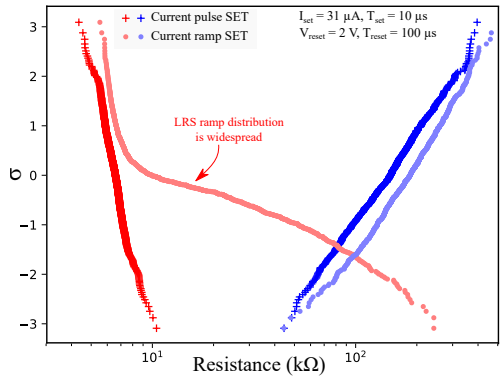


(a) Endurance test, over 500 set/reset cycles. 4 different samples, at set current $I_{set} = 31\mu A$ are shown. The WM looks not degraded over cycling.

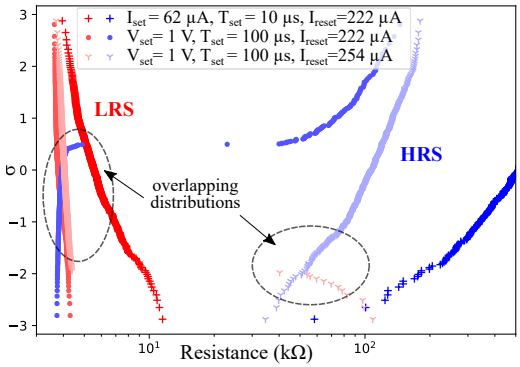


(b) Endurance tests over higher number of cycling: after 2.5k cycles (in gray), and after 8.5k (in red). The distribution tends to shift to the right as the cycle number is increased.

Figure 79 – Endurance tests.



(a) LRS and HRS distributions, resulting from a set process performed by means of a current pulse and a current ramp, and a reset process carried out by a voltage pulse. In the ramp case, the LRS distribution is very widespread and overlaps the HRS distribution.



(b) LRS and HRS distributions, where the reset process was performed through a current pulse of duration $T_{reset} = 10\mu A$ and different amplitudes. Despite the set condition, the resulting dispersions are found to be overlapping, possibly due to excessive (reset) current causing unstable final states.

Figure 80 – Setting with a current ramp and resetting with a current pulse both cause widespread distributions, with overlapping tails.

IMPACT OF A SET CURRENT RAMP After successfully being able to set our samples by means of a current pulse, we proceeded to inspect the impact of a set process performed through a current ramp. In this case, the current is a staircase of initial amplitude $0\mu A$ and final value I_{set}. Figure 80a shows the resulting LRS and HRS distributions obtained for the two cases: a set current pulse and ramp, in order to compare the approaches. It can be seen that, while a current pulse of amplitude $I_{set} = 31\mu A$ can achieve a satisfactory LRS distribution, a ramp of equal (final) amplitude produces a much wider statistical dispersion, and degraded WM.

IMPACT OF A RESET CURRENT PULSE At last, we performed the reset operation by means of a current pulse. As the reset energy cost is higher than the set, we adopted higher current amplitudes. Figure 8ob reports 3 different trials, of various current amplitudes and fixed duration $T_{reset} = 10\mu A$. In all trials, the LRS and HRS distributions end up overlapping, causing reading errors. A possible explanation is that excessive current flows during the reset process, spoiling the final state; in fact, I_{reset} remains constant until the end of the programming pulse, meaning that considerable voltage can drop over a device that reaches its HRS before the end of the programming pulse, possibly causing a fallback to the LRS[6]. As a result, erroneous states and extra damage can arise, making a current-based reset process an unpractical alternative.

4.6 CONCLUSIONS AND PERSPECTIVES

In this chapter, we presented a new circuital solution to perform the programming operation of RRAM. The designed architecture resembles a current DAC, which can be operated both as a programmable voltage and current source. The design specifications, such as the DAC resolution, full-range current and voltage, as well as and linearity requirements, are tuned on the integrated, HfO_2-based, RRAM memory technology.
The first part of the chapter introduced and commented design characteristics, and showed Spice simulation of the main operational blocks. In particular, the proposed architecture benefits of a self write termination mechanism: as soon as the cell enters its LRS, the energy consumed is lessened, as well as the electrical stress over the memory load. A dedicated Write Termination block is also present, in order to save energy after a set process occurs.
In the second part of this chapter, electrical characterization results were exposed. The functionality of the main designed blocks was checked, and we evaluated the circuit potential as a programmable current source. By inspecting the impact of the programming current amplitude and programming time, we concluded that the designed architecture shows good variability and endurance performances, at a reduced energy cost than SOA counterparts. In fact, we were able to reduce the programming current up to a factor of 10: we motivate this result as an interplay of current bias and parasitic capacitance discharge taking place during a set event. Therefore, our circuit revealed as an attractive solution when considering both memory performances and energy expense. Future work can be envisaged, where the architecture is also evaluated as a voltage source and an efficient WT mechanism is integrated to further lessen the energy consumption. Additionally, a Write Verify procedure could be implemented to produce better endurance and Window Margin.

6. In fact, the scenario is different from a reset process performed with a voltage source, where the resistance increase limits the current over the device.

4.7 CHAPITRE 4 - RÉSUMÉ EN FRANÇAIS

Le chapitre 3 a traité de la technique standard de mise en place d'une cellule RRAM, consistant en une impulsion de tension de durée fixe, ainsi que de ses inconvénients majeurs : un fort gaspillage d'énergie et un stress électrique pour la cellule cible. Afin d'atténuer ces effets secondaires, deux nouvelles méthodes de programmation alternatives ont été introduites, où l'opération d'écriture est effectuée au moyen d'un condensateur chargé, d'une inductance ou d'une combinaison des deux. Cependant, ces approches entraînent une empreinte de surface élargie et d'éventuels problèmes de surintensité/surtension lors des transitions de résistance.
Ce chapitre propose une autre solution de conception novatrice, dans laquelle un circuit basé sur le courant DAC est implémenté afin de contrôler à la fois la tension et le courant d'écriture et de limiter la dissipation d'énergie définie.

Le DAC consiste en un réservoir commutable de sources de courant pondérées binaires, qui délivre une amplitude de courant maximale de 127uA à une résolution de 1uA, tandis que la tension de sortie maximale est supérieure à 4V.
La boucle de rétroaction reliant le nœud d'électrode supérieur de la cellule au comparateur de contrôle de polarité du compteur permet de suivre la tension de sortie jusqu'à la référence d'entrée, V_{ref}. Cependant, la sortie est affectée par une ripple de tension en steady-state, dont l'amplitude dépend de la valeur particulière de la résistance de charge RRAM. Son valeur a été estimée dans des plages de tolérance de quelques dizaines de mV. Un compromis entre la complexité de conception et l'amplitude du ripple doit être fait.

La tâche ultime du circuit proposé est de réagir à un événement défini, opportunément couper le biais qui endommage l'échantillon et engendre un gaspillage d'énergie. Le circuit proposé, étant intrinsèquement contrôlé en courant, contrairement aux approches standard basées sur un régulateur de tension linéaire, bénéficie d'un mécanisme d'auto-terminaison. Lorsque la résistance de charge chute, sa tension diminue également, de sorte que la contrainte électrique et les consommations d'énergie sont considérablement réduites dès que la RRAM entre dans son LRS.
Alternativement, le circuit de terminaison d'écriture externe (Write Termination) dédié peut être utilisé pour couper la polarisation à travers la cellule dès qu'une transition de résistance nette se produit. Dans notre implémentation, un filtre passe-haut est utilisé pour déclencher l'état d'un comparateur qui commande le commutateur qui couple le DAC à la cellule mémoire. Il est également envisageable de remplacer la terminaison d'écriture par un protocole de vérification d'écriture (Write Verify), au prix d'un temps de programmation plus long.

Le circuit a été intégré sur une tranche de silicium de 8 pouces, en technologie CMOS 130 nm. Un premier processus de caractérisation électrique des principaux blocs constitutifs est effectué, afin d'évaluer la fiabilité du circuit. Le buffer analogique, l'impédance d'accès et le DAC ont été evalués et validés.
Successivement, nous avons effectué des tests sur des échantillons de RRAM en

faisant fonctionner l'architecture comme une source de courant programmable. En fait, un processus piloté par le courant semble attrayant compte tenu du faible gaspillage d'énergie, même en l'absence d'un mécanisme WT.

En examinant l'impact de l'amplitude du courant de programmation et le temps de programmation, nous avons conclu que l'architecture conçue montre de bonnes performances de variabilité et d'endurance, à un coût énergétique réduit par rapport à ses homologues SOA. En effet, nous avons pu réduire le courant de programmation jusqu'à un facteur 10 : nous motivons ce résultat par une interaction de polarisation de courant et de décharge de capacité parasite se produisant lors d'un événement défini.
Par conséquent, notre circuit s'est révélé être une solution intéressante compte tenu à la fois performances mémoire et dépense énergétique. Des travaux futurs peuvent être envisagés, où l'architecture est également évaluée comme une source de tension et son mécanisme WT est référencé, a fine de réduire encore la consommation d'énergie. De plus, une protocol de Write Verify pourrait être mise en œuvre pour produire une meilleure endurance et un meilleure Window Margin (WM).

SUMMARY AND CONCLUSIONS

This work presented a wide-spread study on the topic of energy efficiency in RRAM arrays. Various approaches were proposed, ranging from device engineering to peripheral circuit design.

CHAPTER 1 gave context to our research. In a market where the need of smaller, faster and low consuming memories is ever growing, RRAM is seen as a major player, thanks to its fast programming time, low write energy and CMOS process compatibility. The general physics behind RRAM working mechanism, as well as SOA design considerations were presented.

CHAPTER 2 introduced the novel idea of RRAM as energy source. Our research is rooted in the fact that that highly resistive devices resemble ionic batteries at the nanoscale. Hence, a disruptive technology can be envisaged, where memory operations are fueled by close-proximity RRAM battery cells, in order to produce highly compact and extremely low power/autonomous devices. In order to benchmark RRAM as energy source, we performed Cyclic Voltammetry tests on a wide range of SOA memory cells. Different materials and geometries were screened, revealing that Ta_2O_5-based samples were the most promising. From electrochemical curves of reduction peaks, we extracted the main features of the technology:

- △ **Local storage:** only a portion of the whole oxide volume is involved in ion conduction, meaning that multiple samples in parallel are a better choice over larger cells to increase energy capacity.
- △ **High energy density** of $3.5pJ\mu m^{-}2$ and power density of $80pW\mu m^{-}2$, which rival with SOA planar supercapacitors.
- △ **Minimum cell area** evaluated was $0.07\mu m^2$, allowing the technology to be extremely compact.

Therefore, the preliminary results presented in this work point in the direction that RRAM as energy source holds great potential, and might be feasible at a more mature stage of research. Further study is demanded in order to fully characterize the technology, where the output voltage per cell, coulombic efficiency and experimental charge/discharge tests are carried out.

CHAPTER 3 tackles the issue of energy waste during RRAM programming by proposing a new circuital solution to perform a set operation, based on a charged capacitor. According to this approach, energy can be stored in a controlled fashion, so that the amount delivered to a target cell is limited. We first proceeded to estimate the size of the programming capacitor with the aid of a circuit model which was both energy and voltage sensitive. We calibrated our model using experimental data of RRAM switching times, obtained when set at a voltage of 1V. We concluded that our approach was, on average, x10 more efficient than the standard CVS method; however, the derived capacitance of 24nF was too large to

be integrated on a chip. Hence, we took advantage of the exponential relationship between RRAM set time and voltage, in order to decrease the energy (capacitance) required for a set process. We achieved proof of concept by means of a discrete-component board, where the smallest programming capacitor amounted to 330pF, when charged at 3V.
Next, we evaluated the viability of an integrated solution, where the parasitic losses can be reduced thanks to the closer proximity of the programming capacitor to the memory. We carried out our analysis considering a SOA 16kb RRAM array, designed in 130nm CMOS technology. We also included parasitic contributions evaluated by Post-Layout-Simulation. Our results were:

- △ **Increasing the charging voltage efficiently reduces the capacitor size.** Hence, technological nodes operating at high Vdd (> 3V) are more compatible with this solution.
- △ **The approach can be integrated on-chip.** A capacitor of 2pF, charged at 4V, is able to deliver enough energy (10pJ) to set a RRAM cell, so that satisfactory performances ($WM = 10$ at 2σ) can be obtained at array level.
- △ **The method should be adapted to the environment.** Larger arrays, being more leaky, might require more than one capacitor; for example, a large array could be divided into n sub-arrays, so that n equally sized capacitors could be used to program all the cells with with satisfactory performances.

Finally, an alternative approach, relying on a charged inductor, was presented. In this situation, the design must consider preventive measures to challenging side effects, such as overvoltages and higher parasitic losses (with respect to the capacitor case). The two methods might also be combined, for example in an architecture which employs a capacitor for the set, and an inductor for the reset process. As perspective, an actual integrated circuit should be fabricated with the quantified parameters, in order experimentally benchmark the proposed approach.

CHAPTER 4 proposed an additional circuital solution to the set energy waste in RRAM, where a current DAC-based architecture was designed, fabricated and tested. The circuit can be operated either as a programmable voltage or current source, and an additional write termination circuitry can be enabled to minimize the energy waste. Due to its current-controlled nature, the system benefits of an intrinsic self-termination, where the voltage across a RRAM undergoing a set process drops as soon as its resistance does. After describing the circuit design, we presented the results of electrical characterization. We were able to validate the functionality of the main constituting blocks, and performed tests on SOA RRAM devices when operating the circuit as a programmable current source. We evaluated the impact of the current amplitude variation and set time duration, over a multitude of samples and cycles. We evaluated our results considering the influence of the metal line parasitic capacitance, concluding that:

- △ **The parasitic capacitance plays a big role** and it influences the overall energy that is delivered to a cell undergoing a set process. Its contribution was found beneficial, and it allows to achieve satisfactory memory performances (variability and endurance) at lower energy consumption: more than a fac-

tor of 10 of improvement could be achieved over other SOA current-driven architectures.

△ **A set current ramp and reset current pulse did not work** withing satisfactory margins. We concluded that these are poor programming strategies, which produced wide, overlapping state distributions.

More extensive characterization can be envisaged in future work. Additionally, the circuit could be tested in voltage-driven mode, with and without the write termination circuitry, in order to inspect the impact on memory performance and energy consumption. More advanced circuit versions can be envisaged, for example operating at sub-microsencods programming times, and integrating additional Write Termination / write Verify systems to minimize consumption and enlarge the Window Margin.

IN CONCLUSION a variety of original strategies have been proposed in this work. Each solution has been formulated for the first time, and thus requires further study, as better detailed in each Chapter's conclusive perspectives, in order to become a practical alternative in standard RRAM arrays. Nevertheless, our results express high potential to both start a brand new application field and improve existing RRAM memory technology.

CONCLUSIONS EN FRANÇAIS

Ce travail a présenté une vaste étude sur le thème de l'efficacité énergétique dans les tableaux RRAM. Diverses approches ont été proposées, allant de l'ingénierie des dispositifs à la conception de circuits périphériques.

LE CHAPITRE 1 a donné un contexte à notre recherche. Dans un marché où le besoin de mémoires plus petites, plus rapides et à faible consommation ne cesse de croître, la RRAM est considérée comme un acteur majeur, grâce à son temps de programmation rapide, sa faible énergie d'écriture et sa compatibilité avec les processus CMOS. La physique générale derrière le mécanisme de travail RRAM, ainsi que les considérations de conception SOA ont été présentées.

LE CHAPITRE 2 introduit l'idée novatrice de la RRAM comme source d'énergie. Notre recherche est ancrée dans le fait que les dispositifs hautement résistifs ressemblent à des batteries ioniques à l'échelle nanométrique. Par conséquent, une technologie de rupture peut être envisagée, où les opérations de mémoire sont alimentées par des cellules de batterie RRAM à proximité, afin de produire des dispositifs très compacts et extrêmement basse consommation/autonomes. Afin de comparer la RRAM comme source d'énergie, nous avons effectué des tests de voltamétrie cyclique sur une large gamme de cellules de mémoire SOA. Différents matériaux et géométries ont été examinés, révélant que les échantillons à base de Ta_2O_5 étaient les plus prometteurs. A partir des courbes électrochimiques des pics de réduction, nous avons extrait les principales caractéristiques de la technologie :

- △ **Stockage local :** seule une partie du volume total d'oxyde est impliquée dans la conduction ionique, ce qui signifie que plusieurs échantillons en parallèle sont un meilleur choix que des cellules plus grandes pour augmenter la capacité énergétique.
- △ **Haute densité d'énergie** de 3,5 $pJ\mu m^{-}2$ et densité de puissance de 80 $pW\mu m^{-}2$, qui rivalisent avec les supercondensateurs planaires SOA.
- △ **Zone de cellule minimale** évaluée était de 0,07 μm^2, permettant à la technologie d'être extrêmement compacte.

Par conséquent, les résultats préliminaires présentés dans ce travail indiquent que la RRAM en tant que source d'énergie présente un grand potentiel et pourrait être réalisable à un stade de recherche plus avancé. Une étude plus approfondie est exigée afin de caractériser complètement la technologie, où la tension de sortie par cellule, l'efficacité coulombique et les tests expérimentaux de charge/décharge sont effectués.

LE CHAPITRE 3 aborde le problème des déchets d'énergie lors de la programmation RRAM en proposant une nouvelle solution circuitique pour effectuer une opération définie, sur la base d'un condensateur chargé. Selon cette approche, l'énergie peut être stockée de manière contrôlée, de sorte que la quantité délivrée à une cellule cible est limitée. Nous avons d'abord procédé à estimer la taille du condensateur de programmation à l'aide d'un modèle de circuit qui était à la fois sensible à l'énergie et à la tension. Nous avons calibré notre modèle en utilisant des

données expérimentales de temps de commutation RRAM, obtenues lorsqu'elles sont définies à une tension de 1V. Nous avons conclu que notre approche était, en moyenne, x10 plus efficace que la méthode CVS standard; Cependant, la capacité dérivée de 24 NF était trop grande pour être intégrée sur une puce. Par conséquent, nous avons profité de la relation exponentielle entre le temps et la tension du RRAM, afin de diminuer l'énergie (capacité) requise pour un processus défini. Nous avons réalisé une preuve de concept au moyen d'un conseil d'administration discret, où le plus petit condensateur de programmation s'élevait à 330 PF, lorsqu'ils sont facturés à 3V.
Ensuite, nous avons évalué la viabilité d'une solution intégrée, où les pertes parasites peuvent être réduites grâce à la proximité plus proche du condensateur de programmation à la mémoire. Nous avons effectué notre analyse compte tenu d'un tableau SOA de 16 Ko RRAM, conçu dans la technologie CMOS 130nm. Nous avons également inclus des contributions parasites évaluées par simulation après la couche. Nos résultats ont été:

- △ **augmentation de la tension de charge réduit efficacement la taille du condensateur.** Par conséquent, les nœuds technologiques fonctionnant à un VDD élevé ($> 3v$) sont plus compatibles avec cette solution.
- △ **L'approche peut être intégrée sur puce.** Un condensateur de 2pf, chargé à 4V, est capable de fournir suffisamment d'énergie (10pj) pour programmer une cellule RRAM, de sorte que des performances satisfaisantes ($WM = 10$ à 2 Sigma) peuvent être obtenues au niveau du tableau.
- △ **La méthode doit être adaptée à l'environnement.** Des tableaux plus grands, étant plus fuites, peuvent nécessiter plus d'un condensateur; Par exemple, un grand tableau pourrait être divisé en sous-arraines n, de sorte que n condensateurs de taille égale puisse être utilisé pour programmer toutes les cellules avec des performances satisfaisantes.

Enfin, une approche alternative, reposant sur une inductance chargée, a été présentée. Dans cette situation, la conception doit tenir compte des mesures préventives des effets secondaires difficiles, tels que des surtensions et des pertes parasitaires plus élevées (en ce qui concerne le cas du condensateur). Les deux méthodes peuvent également être combinées, par exemple dans une architecture qui utilise un condensateur pour l'ensemble et une inductance pour le processus de réinitialisation. En perspective, un circuit intégré réel doit être fabriqué avec les paramètres quantifiés, afin de comparer expérimentalement l'approche proposée.

LE CHAPITRE 4 a proposé une solution circuisée supplémentaire à l'ensemble des déchets d'énergie dans RRAM, où une architecture basée à Current AC DAC a été conçue, fabriquée et testée. Le circuit peut être utilisé soit comme une tension programmable ou une source de courant, et un circuit de terminaison d'écriture supplémentaire peut être activé pour minimiser les déchets d'énergie. En raison de sa nature contrôlée par le courant, le système profite d'une auto-termination intrinsèque, où la tension à travers un RRAM subissant un processus défini baisse dès que sa résistance le fait. Après avoir décrit la conception du circuit, nous avons présenté les résultats de la caractérisation électrique. Nous avons pu valider la fonctionnalité des principaux blocs constitués et effectué des tests sur les disposi-

tifs SOA RRAM lors du fonctionnement du circuit en tant que source de courant programmable. Nous avons évalué l'impact de la variation d'amplitude actuelle et réglé la durée, sur une multitude d'échantillons et de cycles. Nous avons évalué nos résultats compte tenu de l'influence de la capacité parasite de la ligne métallique, concluant que:

△ **La capacité parasite joue un grand rôle** et il influence l'énergie globale qui est livrée à une cellule subissant un processus défini. Sa contribution a été jugée bénéfique et permet d'obtenir des performances de mémoire satisfaisantes (variabilité et endurance) à une consommation d'énergie plus faible: plus d'un facteur de 10 d'amélioration pourrait être obtenu par rapport à d'autres architectures axées sur le courant SOA.

△ **Une rampe de courant définie et une impulsion de courant de réinitialisation ne fonctionnaient pas** dans des marges satisfaisantes. Nous avons conclu que ce sont de mauvaises stratégies de programmation, qui produisaient des distributions d'État larges et chevauchant.

Une caractérisation plus étendue peut être envisagée dans les travaux futurs. De plus, le circuit pourrait être testé en mode basé sur la tension, avec et sans les circuits de terminaison d'écriture, afin d'inspecter l'impact sur les performances de la mémoire et la consommation d'énergie. Des versions de circuits plus avancées peuvent être envisagées, par exemple en fonctionnant dans les temps de programmation des sous-microSencods, et en intégrant des systèmes de terminaison / écriture d'écriture supplémentaires pour minimiser la consommation et agrandir la marge de fenêtre.

EN CONCLUSION Une variété de stratégies originales ont été proposées dans ce travail. Chaque solution a été formulée pour la première fois, et nécessite donc une étude plus approfondie, comme mieux détaillée dans les perspectives concluantes de chaque chapitre, afin de devenir une alternative pratique dans les réseaux RRAM standard. Néanmoins, nos résultats expriment un potentiel élevé à la fois pour démarrer un tout nouveau domaine d'application et améliorer la technologie de mémoire RRAM existante.

APPENDIX

Supplementary material is hereby reported, parted by chapter of reference.

CHAPTER 2

4.7.1 *Reduction peaks during CV tests*

Figure 81 shows 20 CV cycles, performed on a Ta_2O_5-based sample. A current peak appears during negative scan rates, which we associate to a reduction process taking place at the top electrode-electrolyte interface.

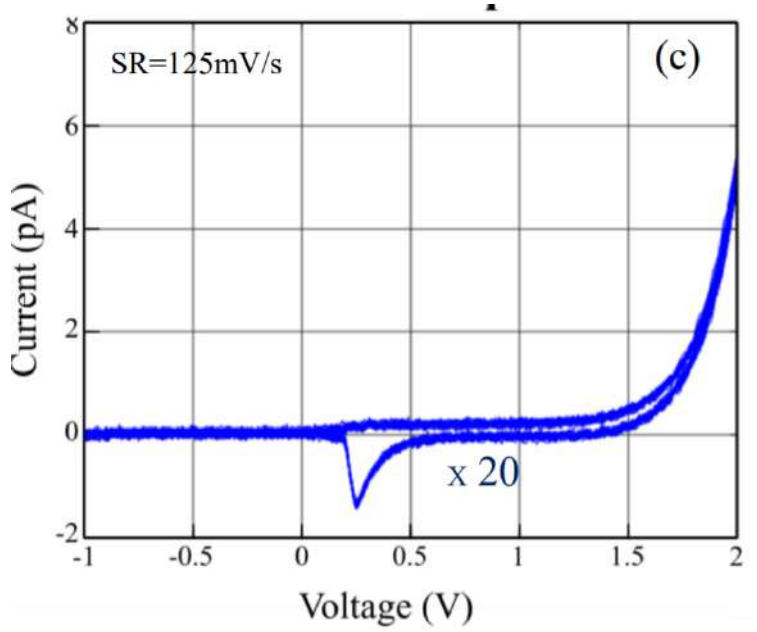


Figure 81 – Current peak appearing during negative sweep rate. 20 CV tests are performed in series, and the resulting plots are over imposed, showing repeatability of the current peak.

4.7.2 *Memory addressing architecture*

Figure 82a and Figure 82b show the implemented decoder's architecture lookup tables. Depending on the configuration of the address bits a_0, a_1, Polarity, voltages V_m and gnd are differently routed to the target top ($t_0 : t_3$) and bottom electrodes ($b_0 : b_3$) of the four 1R-cell RRAMs.

Polarity=0

a_1	a_0	t_0	t_1	t_2	t_3	b_0	b_1	b_2	b_3
0	0	Vm	–	–	–	gnd	–	–	–
0	1	–	Vm	–	–	–	gnd	–	–
1	0	–	–	Vm	–	–	–	gnd	–
1	1	–	–	–	Vm	–	–	–	gnd

(a) Lookup table showing top and bottom electrode values at different $a_0 a_1$ address bits, for polarity flag = 0.

Polarity=1

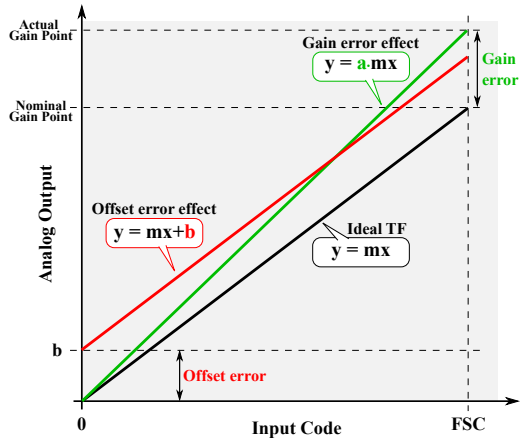
a_1	a_0	t_0	t_1	t_2	t_3	b_0	b_1	b_2	b_3
0	0	gnd	–	–	–	Vm	–	–	–
0	1	–	gnd	–	–	–	Vm	–	–
1	0	–	–	gnd	–	–	–	Vm	–
1	1	–	–	–	gnd	–	–	–	Vm

(b) Lookup table showing top and bottom electrode values at different $a_0 a_1$ address bits, for polarity flag = 1.

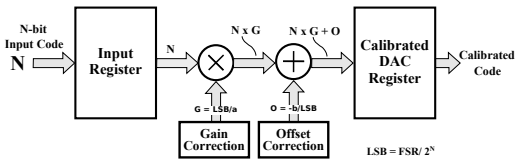
Figure 82 – Memory address block and routing bias.

Offset and Gain errors in DACs

Figure 83a shows offset and gain errors affecting linear DAC transfer functions. The offset error causes vertical shift, while the gain error changes the slope of the output characteristic. Figure Figure 83b shows an example of error correction, which allows to get rid of offset and gain errors: the input code undergoes a multiplication operation, which reestablishes the desired slope (gain correction), and an addition operation, which zeroes the output for zero input code (offset correction).



(a) Linear Transfer Fuctions showing the effect of gain and offset errors, which deviate, respectively, the slope and the origin, with respect to the ideal DAC straight line TF.

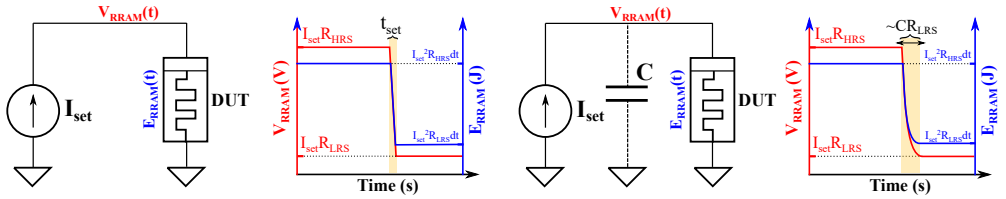


(b) Error calibration, carried out applying gain and offset corrections to the digital input code.

Figure 83 – Gain and offset errors in DACs.

4.7.3 *Impact of the parasitic capacitance on a current-driven set operation*

When the set process is performed through a current source, and assuming the HRS to LRS transition occurs over a narrow time interval, the sudden drop in device resistance causes a sharp voltage drop across its terminals. Similarly, an equally sharp decrease in the instantaneous dissipated energy occurs; this situation is illustrated in Figure 84a. On the other hand, when some parasitic capacitance is present across the sample (in the case of a memory array this contribution can majorly results from the metal line parasitic capacitance) the set time transient is extended, as shown in Figure 84b.



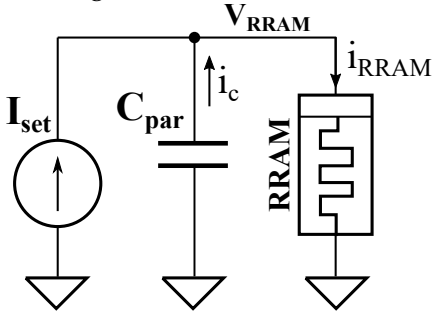
(a) In the absence of parasitics, the voltage across the RRAM DUT, as well as its instantaneous energy, sharply decrease over the set time t_{set}.

(b) In the presence of parasitic capacitance C, the voltage across the DUT (and hence also the energy) are subject to a time delay approximately equal to CR_{LRS}.

Figure 84 – The impact of parasitic capacitance in a current-driven set process.

4.7.4 *Evaluation of the parasitic capacitance energy contribution during a set process*

Given the circuit shown below, the energy dissipated over the RRAM memory cell, during a set process, is calculated. The capacitor is assumed initially charged at the voltage over the RRAM before the set occurs.



I_{set} : set programming current.
i_{RRAM} : current flowing into the memory.
C_{par} : parasitic capacitance.
i_C : current flowing from C_{par}.
t_{set} : time of set.
t_{prog} : programming time.

Voltage over the RRAM before the set event:

$$V_{RRAM,HRS} = I_{set}R_{HRS} \tag{41}$$

Voltage over the RRAM after the set event, until the end of the programming pulse:

$$V_{RRAM,LRS} = I_{set}R_{LRS} \tag{42}$$

The energy stored inside capacitor C_{par}:

$$E_C = \frac{1}{2}C_{par}V^2_{RRAM,HRS} = \frac{1}{2}C_{par}(I_{set}R_{HRS})^2 \tag{43}$$

Energy dissipated over the memory, neglecting the contribution of C_{par}:

$$\begin{aligned} E_{RRAM} &= E_{RRAM,HRS} + E_{RRAM,LRS} \\ &= I_{set} V_{RRAM,HRS} t_{set} + I_{set} V_{RRAM,LRS} (t_{prog} - t_{set}) \\ &= I_{set}^2 R_{HRS} t_{set} + I_{set}^2 R_{LRS} (t_{prog} - t_{set}) \\ &= I_{set}^2 (R_{HRS} t_{set} - R_{LRS} t_{set} + R_{LRS} t_{prog}) \end{aligned} \tag{44}$$

Hence, the fraction of energy delivered by the capacitor, compared to the total:

$$\begin{aligned} \Delta E &= \frac{E_C}{E_C + E_{RRAM}} \\ &= \frac{\frac{1}{2} C_{par} (I_{set} R_{HRS})^2}{\frac{1}{2} C_{par} (I_{set} R_{HRS})^2 + I_{set}^2 (R_{HRS} t_{set} - R_{LRS} t_{set} + R_{LRS} t_{prog})} \\ &= \frac{1}{1 + \frac{2}{C_{par} R_{HRS}^2} (R_{HRS} t_{set} + R_{LRS} (t_{prog} - t_{set})} \\ &= \left[1 + \frac{2}{C_{par}} \left(\frac{t_{set}}{R_{HRS}} + \frac{R_{LRS}}{R_{HRS}^2} (t_{prog} - t_{set}) \right) \right]^{-1} \end{aligned} \tag{45}$$

which, given $R_{HRS} \gg R_{LRS}$ can be approximated as:

$$\Delta E = \left[1 + \frac{2 t_{set}}{C_{par} R_{HRS}} \right]^{-1} \tag{46}$$

It can be seen that:

- $\Delta E \to 1$ if $t_{set} \to 0$: in this case the current source is on for a negligible amount of time, and the RRAM does not dissipate any energy besides the amount delivered by the capacitor.
- $\Delta E \to 0$ if $t_{set} \to$ inf: the capacitor contribution becomes negligible if the set event takes place after a long time.

4.7.5 *Experimental validation of DAC architecture*

Figure 85 shows two DAC transfer functions, as well as static errors, quantified for two different dies. The offset error always resulted negligible, while gain-correction was performed to evaluate non-lineatiry errors. All the samples evaluated were found functional.

4.7.6 *Set tests with DAC in Constant Current Source mode*

This Section reports additional results, obtained when setting RRAM through the DAC in Constant-Current-Source (CCS) mode. Figure 86a and Figure 86b show LRS (in red) and HRS (in blue) normal cumulative distributions (500 samples tested in each). Over our trials, it was confirmed that a programming current of 15µA (at $T_{set} = 10\mu s$) was insufficient to obtain satisfactory distributions, while higher-end current amplitudes stopped being beneficial to the window margin beyond 31µA.

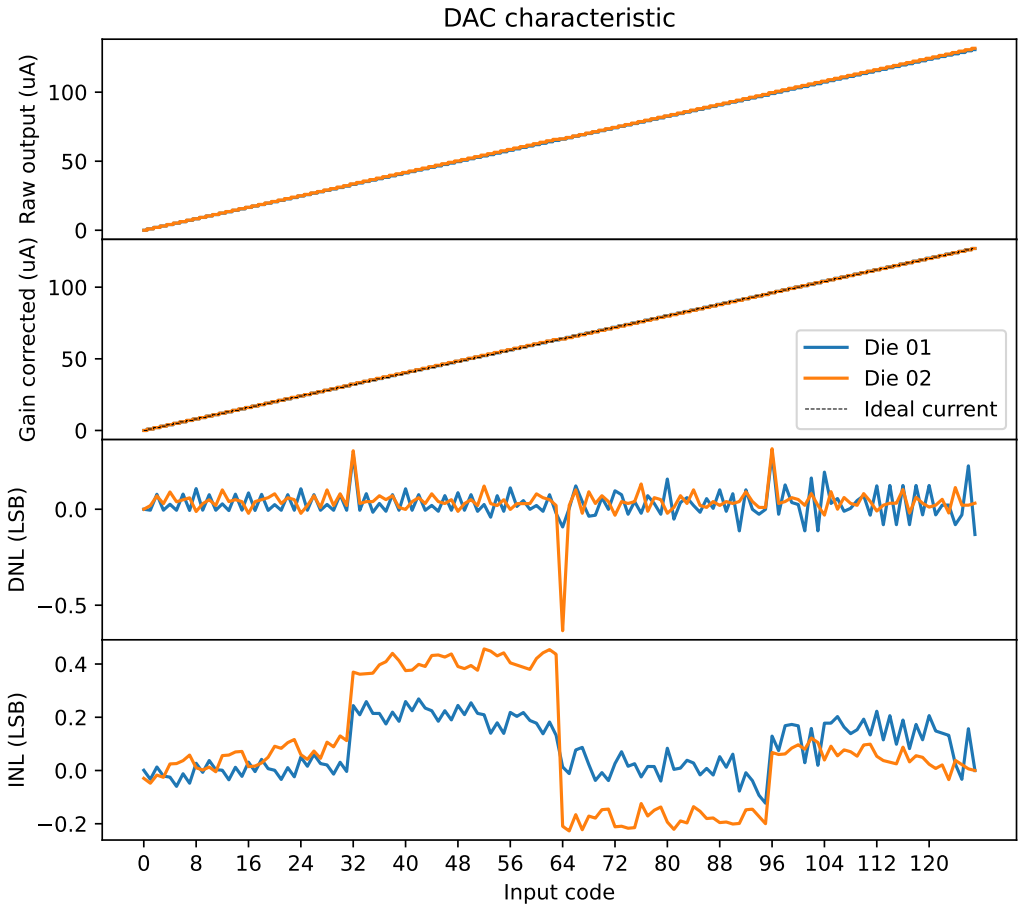
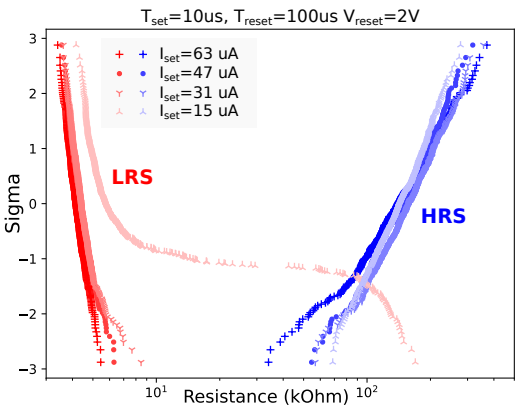
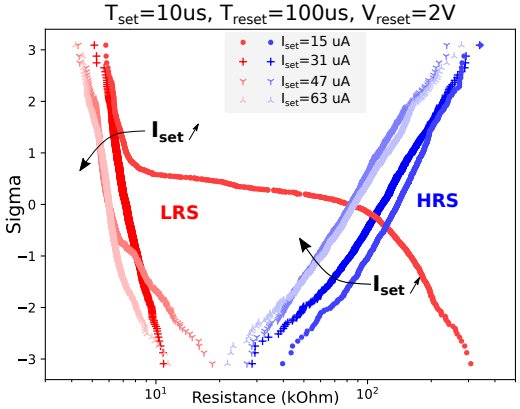


Figure 85 – DAC static errors for two different dies.

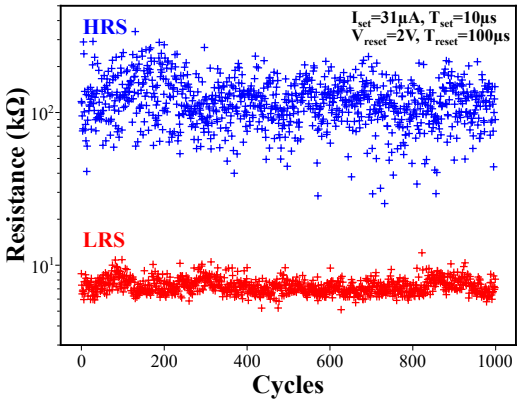
Figure 86c shows an endurance characteristic, obtained setting a RRAM cell with a current pulse of amplitude $I_{set} = 31\mu A$ and duration $T_{set} = 10\mu s$, while resetting with a voltage pulse (amplitude $V_{reset} = 2V$, duration $T_{reset} = 100\mu s$). It can be noted that the LRS and HRS values are never overlapping, up to 10^3 cycles, while their mean values are approximately one order of magnitude apart.



(a) LRS and HRS resistance distributions, overlapping once $I_{set} > 15\mu A$



(b) LRS and HRS distributions. A slight effect of current amplitude can be observed over the distribution dispersion.



(c) Endurance test on a RRAM sample, showing 10^3 set/reset cycles. No resistive state overlap occurs when the set current is equal to 31µA.

Figure 86 – HRS (in blue) and LRS (in red) resistance distributions, obtained ranging the set current I_{set}. As a general trend, the increase of I_{set} initially improves the window margin (tails overlap at $I_{set} = 15\mu A$) but eventually narrows it as the HRS distribution moves to the left.

List of author's patents and publications.

Papers:

(1) Trotti P., Oukassi S., Molas G., Bernard M., Aussenac F., and Pillonnet G. (2021). *In Memory Energy Application for Resistive Random Access Memory. Advanced Electronic Materials*, 7(12), 2100297. https://onlinelibrary.wiley.com/doi/abs/10.1002/aelm.202100297

(2) Trotti P., G. Pillonnet, G. Molas. S. Oukassi and E. Nowak, *Experimental Set-Up For Novel Energy Efficient Charge-based Resistive RAM (RRAM) Switching*. 2020 IEEE 33rd International Conference on Microelectronic Test Structures (ICMTS). IEEE, 2020. https://ieeexplore.ieee.org/abstract/document/9107936

Patents:

(3) Trotti P.., Gabriel M., S. Oukassi and G. Pillonnet, *Energy Recovery In Filamentary Resistive Memories*, U.S. Patent 20210166758A1, Jun. 3, 2021.

(4) A Verdy, G Molas, P. Trotti, A Regev, *Memory comprising a matrix of resistive memory cells, and associated method of interfacing*, US Patent App. 17549162, Jun. 16 2021.

Conference without proceedings:

(5) P. Trotti, G. Molas. S. Oukassi, G. Pillonnet, P. Blaise, M. Bernard and E. Nowak, *Memory and Nanobattery Dual Operation in $Ag/GeS_2/W$ CBRAM.* 2019 Spring Meeting - European Materials Research Society (EMRS), Nice, 2019.

[1] Hiroshi Iwai. « Future of CMOS technology. » In: *2004 Semiconductor Manufacturing Technology Workshop Proceedings (IEEE Cat. No. 04EX846)*. IEEE. 2004, pp. 5–17.

[2] Elie Maricau and Georges Gielen. « CMOS reliability overview. » In: *Analog IC reliability in nanometer CMOS*. Springer, 2013, pp. 15–35.

[3] Larry Zhao. *All About Interconnects*. https://semiengineering.com/all-about-interconnects/. Accessed: 2022-10-06. 2017.

[4] Robert H Havemann and James A Hutchby. « High-performance interconnects: An integration overview. » In: *Proceedings of the IEEE* 89.5 (2001), pp. 586–601.

[5] Pritam Bhattacharjee and Arindam Sadhu. « VLSI Transistor and Interconnect Scaling Overview. » In: *Journal of Electronic Design Technology* 5.1 (2014), pp. 1–15.

[6] W. A. Wulf and S. A. McKee. « Hitting the memory wall: Implications of the obvious. » In: vol. 23. 1995, pp. 20–24.

[7] J. L. Hennessy and D. A. Patterson. *Computer Architecture A Quantitative Approach*. 3nd. San Mateo, CA, USA: Morgan Kaufmann, 2002.

[8] Onur Mutlu, Saugata Ghose, Juan Gómez-Luna, and Rachata Ausavarungnirun. « Processing data where it makes sense: Enabling in-memory computation. » In: *Microprocessors and Microsystems* 67 (2019), pp. 28–41.

[9] Abu Sebastian, Manuel Le Gallo, Riduan Khaddam-Aljameh, and Evangelos Eleftheriou. « Memory devices and applications for in-memory computing. » In: *Nature nanotechnology* 15.7 (2020), pp. 529–544.

[10] Naveen Verma, Hongyang Jia, Hossein Valavi, Yinqi Tang, Murat Ozatay, Lung-Yen Chen, Bonan Zhang, and Peter Deaville. « In-memory computing: Advances and prospects. » In: *IEEE Solid-State Circuits Magazine* 11.3 (2019), pp. 43–55.

[11] Tao Jiang, Qianlong Zhang, Rui Hou, Lin Chai, Sally A Mckee, Zhen Jia, and Ninghui Sun. « Understanding the behavior of in-memory computing workloads. » In: *2014 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE. 2014, pp. 22–30.

[12] Apogeeweb. *In-Memory Computing Technology Overview*. hhttps://www.apogeeweb.net/electron/what-is-in-memory-computing.html. Accessed: 2022-10-06. 2020.

[13] Khaddam-Aljameh R. et al. Sebastian A. Le Gallo M. « Memory devices and applications for in-memory computing. » In: *Nature Nanotechnology* 15 (2020), 529–544. DOI: https://doi.org/10.1038/s41565-020-0655-z. URL: https://www.nature.com/articles/s41565-020-0655-z#citeas.

[14] Gagandeep Singh, Lorenzo Chelini, Stefano Corda, Ahsan Javed Awan, Sander Stuijk, Roel Jordans, Henk Corporaal, and Albert-Jan Boonstra. « A review of near-memory computing architectures: Opportunities and challenges. » In: *2018 21st Euromicro Conference on Digital System Design (DSD)*. IEEE. 2018, pp. 608–617.

[15] Guangdong Zhou, Zhijun Ren, Lidan Wang, Jinggao Wu, Bai Sun, Ankun Zhou, Guanghui Zhang, Shaohui Zheng, Shukai Duan, and Qunliang Song. « Resistive switching memory integrated with amorphous carbon-based nano-generators for self-powered device. » In: *Nano Energy* 63 (2019), p. 103793. ISSN: 2211-2855. DOI: https://doi.org/10.1016/j.nanoen.2019.05.079. URL: https://www.sciencedirect.com/science/article/pii/S2211285519304859.

[16] Ik Kyeong Jin et al. « Self-powered data erasing of nanoscale flash memory by triboelectricity. » In: *Nano Energy* 52 (2018), pp. 63–70. ISSN: 2211-2855. DOI: https://doi.org/10.1016/j.nanoen.2018.07.040. URL: https://www.sciencedirect.com/science/article/pii/S2211285518305287.

[17] Myungjun Kim, Chuljun Lee, Yubin Song, Sang-Mo Koo, Jong-Min Oh, Jiyong Woo, and Daeseok Lee. « Energy-Storing Hybrid 3D Vertical Memory Structure. » In: *IEEE Electron Device Letters* 40.10 (2019), pp. 1622–1625. DOI: 10.1109/LED.2019.2936253.

[18] Yaohan Chu. « Evolution of computer memory structure. » In: *Proceedings of the June 7-10, 1976, national computer conference and exposition*. 1976, pp. 733–748.

[19] David Hemmendinger. *Computer Memory*. https://www.britannica.com/technology/computer-memory. Accessed: 2022-05-31.

[20] EFL. *What Is The Difference Between Primary Memory And Secondary Memory*. https://electronicsforlearning.blogspot.com/2021/04/what-is-difference-between-primary.html. Accessed: 2022-05-31. 2021.

[21] Hai Li and Yiran Chen. « An overview of non-volatile memory technology and the implication for tools and architectures. » In: *2009 Design, Automation & Test in Europe Conference & Exhibition*. IEEE. 2009, pp. 731–736.

[22] Muzaffer A Siddiqi. *Dynamic RAM: technology advancements*. CRC Press, 2017.

[23] Jintao Zhang, Zhuo Wang, and Naveen Verma. « In-memory computation of a machine-learning classifier in a standard 6T SRAM array. » In: *IEEE Journal of Solid-State Circuits* 52.4 (2017), pp. 915–924.

[24] Matthew Marinella. « The future of memory. » In: *2013 IEEE Aerospace Conference*. IEEE. 2013, pp. 1–11.

[25] Jaydeep P Kulkarni, Keejong Kim, and Kaushik Roy. « A 160 mV robust Schmitt trigger based subthreshold SRAM. » In: *IEEE Journal of Solid-State Circuits* 42.10 (2007), pp. 2303–2313.

[26] Sorin Cristoloveanu, KH Lee, MS Parihar, H El Dirani, J Lacord, S Martinie, C Le Royer, J-Ch Barbe, X Mescot, P Fonteneau, et al. « A review of the Z2-FET 1T-DRAM memory: Operation mechanisms and key parameters. » In: *Solid-State Electronics* 143 (2018), pp. 10–19.

[27] Akihiro Nitayama, Yusuke Kohyama, and Katsuhiko Hieda. « Future directions for DRAM memory cell technology. » In: *International Electron Devices Meeting 1998. Technical Digest (Cat. No. 98CH36217)*. IEEE. 1998, pp. 355–358.

[28] Jeff Tyson. *How ROM works*. 2004.

[29] Paolo Pavan, Roberto Bez, Piero Olivo, and Enrico Zanoni. « Flash memory cells-an overview. » In: *Proceedings of the IEEE* 85.8 (1997), pp. 1248–1271.

[30] Roberto Bez, Emilio Camerlenghi, Alberto Modelli, and Angelo Visconti. « Introduction to flash memory. » In: *Proceedings of the IEEE* 91.4 (2003), pp. 489–502.

[31] Mark H Kryder and Chang Soo Kim. « After hard drives—What comes next? » In: *IEEE Transactions on Magnetics* 45.10 (2009), pp. 3406–3413.

[32] IR McFadyen, EE Fullerton, and MJ Carey. « State-of-the-art magnetic hard disk drives. » In: *Mrs Bulletin* 31.5 (2006), pp. 379–383.

[33] John von Neumann. *First draft of a report on the edvac, contract no. w-670-ord-402 moore school of electrical engineering, univ. of penn., philadelphia. Reprinted (in part) in Randell, Brian. 1982. Origins of Digital Computers: Selected Papers*. 1945.

[34] Askquaries. *What Is The von Neumann Bottleneck? It's Meaning and Definition*. https://www.askquaries.com/what-is-the-von-neumann-bottleneck-its-meaning-and-definition/. Accessed: 2022-10-06. 2022.

[35] II Arikpo, FU Ogban, and IE Eteng. « Von neumann architecture and modern computers. » In: *Global Journal of Mathematical Sciences* 6.2 (2007), pp. 97–103.

[36] Guowang Miao, Jens Zander, Ki Won Sung, and Slimane Ben Slimane. *Fundamentals of mobile data networks*. Cambridge University Press, 2016.

[37] Alex Laird. « The Von Neumann architecture topic paper# 3. » In: *Computer Science* 319 (2009), pp. 360–8771.

[38] Philip Jacob, Aamir Zia, Okan Erdogan, Paul Belemjian, Jin-Woo Kim, Michael Chu, R.P. Kraft, J.F. Mcdonald, and Kerry Bernstein. « Mitigating Memory Wall Effects in High-Clock-Rate and Multicore CMOS 3-D Processor Memory Stacks. » In: *Proceedings of the IEEE* 97 (Feb. 2009), pp. 108 –122. DOI: 10.1109/JPROC.2008.2007472.

[39] J Backus. « Can functional programming be liberated from the von Neumann style. » In: *Communications of the ACM* 21.8 (1978), pp. 613–641.

[40] *Persistent Memory Overview*. https://docs.pmem.io/persistent-memory/getting-started-guide/introduction. Accessed: 2021-01-06. 2019.

[41] Donald E Knuth. *TheArtofComputerProgramming: SortingandSearching*. 1973.

[42] Flash Memory Summit. *Persistent Memory*. https://www.flashmemorysummit.com/opt_persistent_memory.html. Accessed: 2022-01-14. 2022.

[43] SNIA Education Committee. *Storage Class Memory – the Future of Solid State Storage*. https://www.snia.org/sites/default/education/tutorials/2009/fall/solid/PhilMills_The_Future_of_Solid_State_Storage.pdf. Accessed: 2022-01-18. 2009.

[44] Tetsuo Endoh, Hiroki Koike, Shoji Ikeda, Takahiro Hanyu, and Hideo Ohno. « An Overview of Nonvolatile Emerging Memories— Spintronics for Working Memories. » In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 6.2 (2016), pp. 109–119. DOI: 10.1109/JETCAS.2016.2547704.

[45] Siva Sivaram. *Storage Class Memory: Can the Promise be Fulfilled?* https://blog.westerndigital.com/storage-class-memory-3d-nand-lessons/. Accessed: 2022-01-14. 2016.

[46] W.I. Kinney, W. Shepherd, W. Miller, J. Evans, and R. Womack. « A non-volatile memory cell based on ferroelectric storage capacitors. » In: *1987 International Electron Devices Meeting*. 1987, pp. 850–851. DOI: 10.1109/IEDM.1987.191567.

[47] N. Setter et al. « Ferroelectric thin films: Review of materials, properties, and applications. » In: *Journal of Applied Physics* 100.5 (2006), p. 051606. DOI: 10.1063/1.2336999. URL: https://doi.org/10.1063/1.2336999.

[48] Electronics-Notes. *What is FRAM memory?: ferroelectric RAM*. https://www.electronics-notes.com/articles/electronic_components/semiconductor-ic-memory/fram-ferroelectric-ram-memory.php. Accessed: 2022-01-19.

[49] Wenxiu Gao, Yi Zhu, Yaojin Wang, Guoliang Yuan, and Jun-Ming Liu. « A review of flexible perovskite oxide ferroelectric films and their application. » In: *Journal of Materiomics* 6.1 (2020), pp. 1–16.

[50] T Mikolajick, S Slesazeck, H Mulaosmanovic, MH Park, S Fichtner, PD Lomenzo, M Hoffmann, and U Schroeder. « Next generation ferroelectric materials for semiconductor process integration and their applications. » In: *Journal of Applied Physics* 129.10 (2021), p. 100901.

[51] T Francois, J Coignus, A Makosiej, B Giraud, C Carabasse, J Barbot, S Martin, N Castellani, T Magis, H Grampeix, et al. « 16kbit HfO 2: Si-based 1T-1C FeRAM Arrays Demonstrating High Performance Operation and Solder Reflow Compatibility. » In: *2021 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2021, pp. 33–1.

[52] Masaharu Kobayashi, Nozomu Ueyama, Kyungmin Jang, and Toshiro Hiramoto. « Experimental Demonstration of a Nonvolatile SRAM With Ferroelectric HfO₂ Capacitor for Normally Off Application. » In: *IEEE Journal of the Electron Devices Society* 6 (2018), pp. 280–285. DOI: 10.1109/JEDS.2018.2800090.

[53] E. Chen, D. Lottis, A. Driskill-Smith, D. Druist, V. Nikitin, S. Watts, X. Tang, and D. Apalkov. « Non-volatile spin-transfer torque RAM (STT-RAM). » In: *68th Device Research Conference*. 2010, pp. 249–252. DOI: 10.1109/DRC.2010.5551975.

[54] Emre Kültürsay, Mahmut Kandemir, Anand Sivasubramaniam, and Onur Mutlu. « Evaluating STT-RAM as an energy-efficient main memory alternative. » In: *2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. 2013, pp. 256–267. DOI: 10.1109/ISPASS.2013.6557176.

[55] Sarath Mohanachandran Nair et al. « Workload-Aware Electromigration Analysis in Emerging Spintronic Memory Arrays. » In: *IEEE Transactions on Device and Materials Reliability* 21.2 (2021), pp. 258–266. DOI: 10.1109/TDMR.2021.3074251.

[56] Stanford R. Ovshinsky. « Reversible Electrical Switching Phenomena in Disordered Structures. » In: *Phys. Rev. Lett.* 21 (20 1968), pp. 1450–1453. DOI: 10.1103/PhysRevLett.21.1450. URL: https://link.aps.org/doi/10.1103/PhysRevLett.21.1450.

[57] RE Simpson, M Krbal, P Fons, AV Kolobov, J Tominaga, T Uruga, and H Tanida. « Toward the ultimate limit of phase change in Ge2Sb2Te5. » In: *Nano letters* 10.2 (2010), pp. 414–419.

[58] Manan Suri, Daniele Garbin, Olivier Bichler, Damien Querlioz, Dominique Vuillaume, Christian Gamrat, and Barbara DeSalvo. « Impact of PCM resistance-drift in neuromorphic systems and drift-mitigation strategy. » In: *2013 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*. IEEE. 2013, pp. 140–145.

[59] PH Nielsen and NM Bashara. « The reversible voltage-induced initial resistance in the negative resistance sandwich structure. » In: *IEEE Transactions on Electron Devices* 11.5 (1964), pp. 243–244.

[60] H.-S. Philip Wong, Heng-Yuan Lee, Shimeng Yu, Yu-Sheng Chen, Yi Wu, Pang-Shiu Chen, Byoungil Lee, Frederick T. Chen, and Ming-Jinn Tsai. « Metal–Oxide RRAM. » In: *Proceedings of the IEEE* 100.6 (2012), pp. 1951–1970. DOI: 10.1109/JPROC.2012.2190369.

[61] H.-S. Philip Wong, Heng-Yuan Lee, Shimeng Yu, Yu-Sheng Chen, Yi Wu, Pang-Shiu Chen, Byoungil Lee, Frederick T. Chen, and Ming-Jinn Tsai. « Metal–Oxide RRAM. » In: *Proceedings of the IEEE* 100.6 (2012), pp. 1951–1970. DOI: 10.1109/JPROC.2012.2190369.

[62] TW Hickmott. « Low-frequency negative resistance in thin anodic oxide films. » In: *Journal of Applied Physics* 33.9 (1962), pp. 2669–2682.

[63] G Dearnaley, AM Stoneham, and DV Morgan. « Electrical phenomena in amorphous oxide films. » In: *Reports on Progress in Physics* 33.3 (1970), p. 1129.

[64] JG Simmons. « Conduction in thin dielectric films. » In: *Journal of Physics D: Applied Physics* 4.5 (1971), p. 613.

[65] IG Baek, MS Lee, S Seo, MJ Lee, DH Seo, D-S Suh, JC Park, SO Park, HS Kim, IK Yoo, et al. « Highly scalable nonvolatile resistive memory using simple binary oxide driven by asymmetric unipolar voltage pulses. » In: *IEDM Technical Digest. IEEE International Electron Devices Meeting, 2004.* IEEE. 2004, pp. 587–590.

[66] Rainer Waser, Regina Dittmann, Georgi Staikov, and Kristof Szot. « Redox-based resistive switching memories–nanoionic mechanisms, prospects, and challenges. » In: *Advanced materials* 21.25-26 (2009), pp. 2632–2663.

[67] Marc Bocquet, Tifenn Hirztlin, J-O Klein, Etienne Nowak, Elisa Vianello, J-M Portal, and Damien Querlioz. « In-memory and error-immune differential RRAM implementation of binarized deep neural networks. » In: *2018 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2018, pp. 20–6.

[68] Furqan Zahoor, Tun Zainal Azni Zulkifli, and Farooq Ahmad Khanday. « Resistive random access memory (RRAM): an overview of materials, switching mechanism, performance, multilevel cell (MLC) storage, modeling, and applications. » In: *Nanoscale research letters* 15.1 (2020), pp. 1–26.

[69] Shimeng Yu, Byoungil Lee, and H-S Philip Wong. « Metal oxide resistive switching memory. » In: *Functional Metal Oxide Nanostructures* (2012), pp. 303–335.

[70] H.-S. Philip Wong, Heng-Yuan Lee, Shimeng Yu, Yu-Sheng Chen, Yi Wu, Pang-Shiu Chen, Byoungil Lee, Frederick T. Chen, and Ming-Jinn Tsai. « Metal–Oxide RRAM. » In: *Proceedings of the IEEE* 100.6 (2012), pp. 1951–1970. DOI: 10.1109/JPROC.2012.2190369.

[71] Paola Trotti, Sami Oukassi, Gabriel Molas, Mathieu Bernard, François Aussenac, and Gaël Pillonnet. « In Memory Energy Application for Resistive Random Access Memory. » In: *Advanced Electronic Materials* 7.12 (2021), p. 2100297.

[72] C Nail, G Molas, P Blaise, G Piccolboni, B Sklenard, C Cagli, M Bernard, A Roule, M Azzaz, E Vianello, et al. « Understanding RRAM endurance, retention and window margin trade-off using experimental results and simulations. » In: *2016 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2016, pp. 4–5.

[73] Writam Banerjee. « Challenges and applications of emerging nonvolatile memory devices. » In: *Electronics* 9.6 (2020), p. 1029.

[74] Peyman Pouyan, Esteve Amat, Said Hamdioui, and Antonio Rubio. « RRAM variability and its mitigation schemes. » In: *2016 26th international workshop on power and timing modeling, optimization and simulation (PATMOS)*. IEEE. 2016, pp. 141–146.

[75] Gabriel Molas, Gilbert Sassine, Cecile Nail, Diego Alfaro Robayo, Jean-François Nodin, Carlo Cagli, Jean Coignus, Philippe Blaise, and Etienne Nowak. « Resistive memories (RRAM) variability: challenges and solutions. » In: *ECS Transactions* 86.3 (2018), p. 35.

[76] Ximeng Guan, Shimeng Yu, and HS Philip Wong. « On the variability of HfOx RRAM: From numerical simulation to compact modeling. » In: *Nanotechnology 2012: Electronics, Devices, Fabrication, MEMS, Fluidics and Computational-2012 NSTI Nanotechnology Conference and Expo, NSTI-Nanotech 2012*. 2012, pp. 815–820.

[77] Alessandro Grossi, Damian Walczyk, Cristian Zambelli, Enrique Miranda, Piero Olivo, Valeriy Stikanov, Alessandro Feriani, Jordi Sune, Gunter Schoof, Rolf Kraemer, et al. « Impact of intercell and intracell variability on forming and switching parameters in RRAM arrays. » In: *IEEE Transactions on Electron Devices* 62.8 (2015), pp. 2502–2509.

[78] Yun-Feng Kao, Wei Cheng Zhuang, Chrong-Jung Lin, and Ya-Chin King. « A study of the variability in contact resistive random access memory by stochastic vacancy model. » In: *Nanoscale research letters* 13.1 (2018), pp. 1–10.

[79] P. Trotti, G. Pillonet, G. Molas, S. Oukassi, and E. Nowak. « Experimental Set-Up For Novel Energy Efficient Charge-based Resistive RAM (RRAM) Switching. » In: *2020 IEEE 33rd International Conference on Microelectronic Test Structures (ICMTS)*. 2020, pp. 1–5. DOI: 10.1109/ICMTS48187.2020.9107936.

[80] Peyman Pouyan, Esteve Amat, Said Hamdioui, and Antonio Rubio. « RRAM variability and its mitigation schemes. » In: *2016 26th international workshop on power and timing modeling, optimization and simulation (PATMOS)*. IEEE. 2016, pp. 141–146.

[81] D Walczyk, Ch Walczyk, T Schroeder, T Bertaud, M Sowińska, M Lukosius, M Fraschke, B Tillack, and Ch Wenger. « Resistive switching characteristics of CMOS embedded HfO2-based 1T1R cells. » In: *Microelectronic Engineering* 88.7 (2011), pp. 1133–1135.

[82] Huaqiang Wu, Yan Liao, Bin Gao, Debanjan Jana, and He Qian. « RRAM Cross-Point Arrays. » In: *3D Flash Memories*. Springer, 2016, pp. 223–260.

[83] Yilkal A Belay, A Cabrini, and G Torelli. « Analysis of array biasing in cross-point memories for leakage power minimization. » In: *2017 13th Conference on Ph. D. Research in Microelectronics and Electronics (PRIME)*. IEEE. 2017, pp. 17–20.

[84] Taehoon Kim, Hyejung Choi, Myoungsub Kim, Jaeyun Yi, Donghoon Kim, Sunglae Cho, Hyunmin Lee, Changyoun Hwang, Eung-Rim Hwang, Jeongho Song, et al. « High-performance, cost-effective 2z nm two-deck cross-point memory integrated by self-align scheme for 128 Gb SCM. » In: *2018 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2018, pp. 37–1.

[85] Min Zhu, Kun Ren, and Zhitang Song. « Ovonic threshold switching selectors for three-dimensional stackable phase-change memory. » In: *MRS Bulletin* 44.9 (2019), pp. 715–720.

[86] HY Cheng, WC Chien, IT Kuo, CW Yeh, L Gignac, W Kim, EK Lai, YF Lin, RL Bruce, C Lavoie, et al. « Ultra-high endurance and low I OFF selector based on AsSeGe chalcogenides for wide memory window 3D stackable crosspoint memory. » In: *2018 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2018, pp. 37–3.

[87] HY Cheng, WC Chien, IT Kuo, EK Lai, Y Zhu, JL Jordan-Sweet, A Ray, F Carta, FM Lee, PH Tseng, et al. « An ultra high endurance and thermally stable selector based on TeAsGeSiSe chalcogenides compatible with BEOL IC Integration for cross-point PCM. » In: *2017 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2017, pp. 2–2.

[88] Anthonin Verdy, Francesco d'Acapito, Jean-Baptiste Dory, Gabriele Navarro, Mathieu Bernard, and Pierre Noé. « Effect of Nitrogen on the Amorphous Structure and Subthreshold Electrical Conduction of GeSeSb-Based Ovonic Threshold Switching Thin Films. » In: *physica status solidi (RRL)–Rapid Research Letters* 14.5 (2020), p. 1900548.

[89] Wolfgang Arden, Michel Brillouët, Patrick Cogez, Mart Graef, Bert Huizing, and R. Mahnkopf. « " More-than-Moore " White Paper. » In: 2010.

[90] Lucanos Strambini et al. « Three-dimensional silicon-integrated capacitor with unprecedented areal capacitance for on-chip energy storage. » In: *Nano Energy* 68 (2020), p. 104281. ISSN: 2211-2855. DOI: https://doi.org/10.1016/j.nanoen.2019.104281. URL: https://www.sciencedirect.com/science/article/pii/S2211285519309887.

[91] Valentin Sallaz, Sami Oukassi, Frédéric Voiron, Raphaël Salot, and David Berardan. « Assessing the potential of LiPON-based electrical double layer microsupercapacitors for on-chip power storage. » In: *Journal of Power Sources* 451 (2020), p. 227786. ISSN: 0378-7753. DOI: https://doi.org/10.1016/j.jpowsour.2020.227786. URL: https://www.sciencedirect.com/science/article/pii/S0378775320300896.

[92] S Oukassi, A Bazin, C Secouard, I Chevalier, S Poncet, S Poulet, JM Boissel, F Geffraye, J Brun, and R Salot. « Millimeter scale thin film batteries for integrated high energy density storage. » In: *2019 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2019, pp. 26–1.

[93] Chen Wang, Jingkun Mao, Giuseppe Selli, Shaofeng Luan, Lin Zhang, Jun Fan, David J Pommerenke, Richard E DuBroff, and James L Drewniak. « An efficient approach for power delivery network design with closed-form expressions for parasitic interconnect inductances. » In: *IEEE Transactions on Advanced Packaging* 29.2 (2006), pp. 320–334.

[94] Kirk Prall. « Benchmarking and metrics for emerging memory. » In: *2017 IEEE International Memory Workshop (IMW)*. IEEE. 2017, pp. 1–5.

[95] Jeongsup Lee et al. « A Self-Tuning IoT Processor Using Leakage-Ratio Measurement for Energy-Optimal Operation. » In: *IEEE Journal of Solid-State Circuits* 55.1 (2020), pp. 87–97. DOI: 10.1109/JSSC.2019.2939890.

[96] Paul A Merolla, John V Arthur, Rodrigo Alvarez-Icaza, Andrew S Cassidy, Jun Sawada, Filipp Akopyan, Bryan L Jackson, Nabil Imam, Chen Guo, Yutaka Nakamura, et al. « A million spiking-neuron integrated circuit with a scalable communication network and interface. » In: *Science* 345.6197 (2014), pp. 668–673.

[97] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. « Loihi: A neuromorphic manycore processor with on-chip learning. » In: *Ieee Micro* 38.1 (2018), pp. 82–99.

[98] Reiji Mochida, Kazuyuki Kouno, Yuriko Hayata, Masayoshi Nakayama, Takashi Ono, Hitoshi Suwa, Ryutaro Yasuhara, Koji Katayama, Takumi Mikawa, and Yasushi Gohou. « A 4M synapses integrated analog ReRAM based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture. » In: *2018 IEEE Symposium on VLSI Technology*. IEEE. 2018, pp. 175–176.

[99] A Valentian, F Rummens, E Vianello, T Mesquida, C Lecat-Mathieu de Boissac, O Bichler, and C Reita. « Fully integrated spiking neural network with analog neurons and RRAM synapses. » In: *2019 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2019, pp. 14–3.

[100] I. Valov, E. Linn, S. Tappertzhofen, J. Schmelzer S.and Van den Hurk, F. Lentz, and R. Waser. « Nanobatteries in redox-based resistive switches require extension of memristor theory. » In: *Nature Communications* 4 (2013), pp. 2041–1723. DOI: 10.1038/ncomms2784.

[101] Rainer Waser, Regina Dittmann, Georgi Staikov, and Kristof Szot. « Redox-Based Resistive Switching Memories – Nanoionic Mechanisms, Prospects, and Challenges. » In: *Advanced Materials* 21.25-26 (2009), pp. 2632–2663. DOI: https://doi.org/10.1002/adma.200900375.

[102] C. Schindler, G. Staikov, and R. Waser. « Electrode kinetics of Cu–SiO2-based resistive switching cells: Overcoming the voltage-time dilemma of electrochemical metallization memories. » In: *Applied Physics Letters* 94.7 (2009), p. 072109. DOI: 10.1063/1.3077310.

[103] Hong-Yu Chen, Stefano Brivio, Che-Chia Chang, Jacopo Frascaroli, Tuo-Hung Hou, Boris Hudec, Ming Liu, Hangbing Lv, Gabriel Molas, Joon Sohn, et al. « Resistive random access memory (RRAM) technology: From material, device, selector, 3D integration to bottom-up fabrication. » In: *Journal of Electroceramics* 39.1 (2017), pp. 21–38.

[104] Tong Liu, Yuhong Kang, Sarah El-Helw, Tanmay Potnis, and Marius Orlowski. « Physics of the voltage constant in multilevel switching of conductive bridge resistive memory. » In: *Japanese Journal of Applied Physics* 52.8R (2013), p. 084202.

[105] Tohru Tsuruoka, Ilia Valov, Stefan Tappertzhofen, Jan van den Hurk, Tsuyoshi Hasegawa, Rainer Waser, and Masakazu Aono. « Redox reactions at Cu, Ag/Ta2O5 interfaces and the effects of Ta2O5 film density on the forming process in atomic switch structures. » In: *Advanced functional materials* 25.40 (2015), pp. 6374–6381.

[106] Piero Zanello, Carlo Nervi, and Fabrizia Fabrizi De Biani. *Inorganic electrochemistry: theory, practice and application*. Royal Society of Chemistry, 2019.

[107] Gary A Mabbott. « An introduction to cyclic voltammetry. » In: *Journal of Chemical education* 60.9 (1983), p. 697.

[108] Peter T Kissinger and William R Heineman. « Cyclic voltammetry. » In: *Journal of chemical education* 60.9 (1983), p. 702.

[109] Stefan Tappertzhofen, Hans Mündelein, Ilia Valov, and Rainer Waser. « Nanoionic transport and electrochemical reactions in resistively switching silicon dioxide. » In: *Nanoscale* 4.10 (2012), pp. 3040–3043.

[110] B Lovrecek, I Mekjavic, FM Metikos-Hukovic, AJ Bard, R Parsons, and J Jordan. « Standard Potentials in Aqueous Solution. » In: *Bismuth, Marcel Dekker, Inc, NY and Basel* (1985).

[111] Carlo Cagli, Gabriel Molas, Michel Harrand, Sophie Bernasconi, Christelle Charpin, Khalil El Hajjam, Jean Francois Nodin, and Gilles Reimbold. « Study of the Energy Consumption Optimization on RRAM Memory Array for SCM Applications. » In: *2017 IEEE International Memory Workshop (IMW)*. 2017, pp. 1–4. DOI: 10.1109/IMW.2017.7939106.

[112] Dake Wang. In: 52.6 (2017), p. 065019. DOI: 10.1088/1361-6552/aa8973. URL: https://doi.org/10.1088/1361-6552/aa8973.

[113] Sami M Al-Jaber, Iyad Saadeddin, et al. « Theoretical and Experimental Analysis of Energy in Charging a Capacitor by Step-Wise Potential. » In: *Journal of Applied Mathematics and Physics* 8.01 (2019), p. 38.

[114] Graham P Boechler, Jean M Whitney, Craig S Lent, Alexei O Orlov, and Gregory L Snider. « Fundamental limits of energy dissipation in charge-based computing. » In: *Applied Physics Letters* 97.10 (2010), p. 103502.

[115] Zhang et al. « Programming resistive switching memory by a charged capacitor. » In: *Appl. Phys. A* 102 (2011), 1003–1007. DOI: 10.1007/s00339-011-6320-5.

[116] Gilbert Sassine et al. « Sub-pJ consumption and short latency time in RRAM arrays for high endurance applications. » In: *2018 IEEE International Reliability Physics Symposium (IRPS)*. 2018, P–MY.2–1–P–MY.2–5. DOI: 10.1109/IRPS.2018.8353675.

[117] « Reliability screening of high-k dielectrics based on voltage ramp stress. » In: *Microelectronics Reliability* 47.4 (2007). 14th Workshop on Dielectrics in Microelectronics (WoDiM 2006), pp. 513–517. ISSN: 0026-2714. DOI: https://doi.org/10.1016/j.microrel.2007.01.030.

[118] E.Y. Wu, A. Vayshenker, E. Nowak, J. Sune, R.-P. Vollertsen, W. Lai, and D. Harmon. « Experimental evidence of T/sub BD/ power-law for voltage dependence of oxide breakdown in ultrathin gate oxides. » In: *IEEE Transactions on Electron Devices* 49.12 (2002), pp. 2244–2253. DOI: 10.1109/TED.2002.805606.

[119] Moritz von Witzleben, Stefan Wiefels, Andreas Kindsmuller, Pascal Stasner, Fenja Berg, Felix Cuppers, Susanne Hoffmann-Eifert, Rainer Waser, Stephan Menzel, and Ulrich Bottger. « Intrinsic RESET speed limit of valence change memories. » In: *ACS Applied Electronic Materials* 3.12 (2021), pp. 5563–5572.

[120] Ulrich Böttger, Moritz von Witzleben, Viktor Havel, Karsten Fleck, Vikas Rana, Rainer Waser, and Stephan Menzel. « Picosecond multilevel resistive switching in tantalum oxide thin films. » In: *Scientific reports* 10.1 (2020), pp. 1–9.

[121] Gilbert Sassine et al. « Sub-pJ consumption and short latency time in RRAM arrays for high endurance applications. » In: *2018 IEEE International Reliability Physics Symposium (IRPS)*. 2018, P–MY.2–1–P–MY.2–5. DOI: 10.1109/IRPS.2018.8353675.

[122] Gilbert Sassine, Carlo Cagli, Jean-François Nodin, Gabriel Molas, and Etienne Nowak. « Novel computing method for short programming time and low energy consumption in HfO 2 based RRAM arrays. » In: *IEEE Journal of the Electron Devices Society* 6 (2018), pp. 696–702.

[123] Alexandre Levisse, Marc Bocquet, Marco Rios, Mouhamad Alayan, Mathieu Moreau, Etienne Nowak, Gabriel Molas, Elisa Vianello, David Atienza, and Jean-Michel Portal. « Write termination circuits for RRAM: A holistic approach from technology to application considerations. » In: *Ieee Access* 8 (2020), pp. 109297–109308.

[124] Ping Zhou, Bo Zhao, Jun Yang, and Youtao Zhang. « Energy reduction for STT-RAM using early write termination. » In: *2009 IEEE/ACM International Conference on Computer-Aided Design-Digest of Technical Papers*. IEEE. 2009, pp. 264–268.

[125] Hassen Aziza, Karine Coulié, and Wenceslas Rahajandraibe. « Design Considerations Towards Zero-Variability Resistive RAMs in HRS State. » In: *2021 IEEE 22nd Latin American Test Symposium (LATS)*. IEEE. 2021, pp. 1–5.

[126] Michael Day. « Understanding low drop out (LDO) regulators. » In: *Texas Instruments* 16 (2006).

[127] Analog Devices. *The Fundamentals of LDO Design and Applications*. https://www.analog.com/en/design-center/landing-pages/001/fundamentals-of-ldo-design-and-applications.html. Accessed: 2022-06-17.

[128] ADP1740 REVERSE. « The Fundamentals of LDO Design and Application. » In: ().

[129] Behzad Razavi. « The current-steering DAC [a circuit for all seasons]. » In: *IEEE Solid-State Circuits Magazine* 10.1 (2018), pp. 11–15.

[130] BenKenton T. Veeder. *Digital Converters for Image Sensors*. SPIE, 2015.

[131] « Current Steering DACs. » In: *Wide-Bandwidth High-Dynamic Range D/A Converters*. Boston, MA: Springer US, 2006, pp. 25–30. ISBN: 978-0-387-30416-8. DOI: 10.1007/0-387-30416-9_03. URL: https://doi.org/10.1007/0-387-30416-9_03.

[132] Texas Instruments. « Understanding data converters. » In: *Application report* (1995).

[133] Fernando Leonel Aguirre, Sebastián Matías Pazos, Félix Palumbo, Jordi Suñé, and Enrique Miranda. « Application of the Quasi-Static Memdiode Model in Cross-Point Arrays for Large Dataset Pattern Recognition. » In: *IEEE Access* 8 (2020), pp. 202174–202193.

[134] Wonbo Shim, Jae-sun Seo, and Shimeng Yu. « Two-step write–verify scheme and impact of the read noise in multilevel RRAM-based inference engine. » In: *Semiconductor Science and Technology* 35.11 (2020), p. 115026.

[135] Shimeng Yu, Ximeng Guan, and H.-S. Philip Wong. « On the Switching Parameter Variation of Metal Oxide RRAM—Part II: Model Corroboration and Device Design Strategy. » In: *IEEE Transactions on Electron Devices* 59.4 (2012), pp. 1183–1188. DOI: 10.1109/TED.2012.2184544.

[136] David M Nminibapiel, Dmitry Veksler, Pragya R Shrestha, Jason P Campbell, Jason T Ryan, Helmut Baumgart, and Kin P Cheung. « Impact of RRAM read fluctuations on the program-verify approach. » In: *IEEE Electron Device Letters* 38.6 (2017), pp. 736–739.

[137] Hassan Aziza, Said Hamdioui, Moritz Fieback, Mottaqiallah Taouil, Mathieu Moreau, Patrick Girard, Arnaud Virazel, and Karine Coulié. « Multi-level control of resistive ram (Rram) using a write termination to achieve 4 bits/cell in high resistance state. » In: *Electronics* 10.18 (2021), p. 2222.

[138] Andrea Fantini, Georgi Gorine, Robin Degraeve, Ludovic Goux, Chao-Yang Chen, Augusto Redolfi, Sergiu Clima, Alessandro Cabrini, Guido Torelli, and Malgorzata Jurczak. « Intrinsic program instability in HfO2 RRAM and consequences on program algorithms. » In: *2015 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2015, pp. 7–5.

[139] Texas Instruments. *Precision DACs: DC Specifications*. https://training.ti.com/precision-dacs-dc-specifications. Accessed: 2022-07-07. 2015.

[140] Janet Heath. *Compensating for DAC Offset and Gain Error*. https://www.analogictips.com/compensating-dac-offset-gain-error. Accessed: 2022-07-07. 2017.

[141] David Fry. *Adjust and Calibrate Offset/Gain Error in Precision DAC Calculation*. https://www.maximintegrated.com/en/design/technical-documents/tutorials/4/4602.html. Accessed: 2022-07-07. 2009.

[142] S Seo, MJ Lee, DH Seo, EJ Jeoung, D-S Suh, YS Joung, IK Yoo, IR Hwang, SH Kim, IS Byun, et al. « Reproducible resistance switching in polycrystalline NiO films. » In: *Applied Physics Letters* 85.23 (2004), pp. 5655–5657.

[143] Pulkit Jain, Umut Arslan, Meenakshi Sekhar, Blake C Lin, Liqiong Wei, Tanaya Sahu, Juan Alzate-Vinasco, Ajay Vangapaty, Mesut Meterelliyoz, Nathan Strutt, et al. « 13.2 A 3.6 Mb 10.1 Mb/mm 2 embedded non-volatile ReRAM macro in 22nm FinFET technology with adaptive forming/set/reset schemes yielding down to 0.5 V with sensing time of 5ns at 0.7 V. » In: *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE. 2019, pp. 212–214.

[144] Chung-Cheng Chou, Zheng-Jun Lin, Pei-Ling Tseng, Chih-Feng Li, Chih-Yang Chang, Wei-Chi Chen, Yu-Der Chih, and Tsung-Yung Jonathan Chang. « An N40 256K× 44 embedded RRAM macro with SL-precharge SA and low-voltage current limiter to improve read and write performance. » In: *2018 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE. 2018, pp. 478–480.