Physical Simulation of the ReRAM reliability and retention through Kinetic Monte Carlo modeling

Master's degree in Nanotechnologies for Smart and Integrated Systems

A.Y. 2023/2024

Supervisors:
Prof. Carlo Ricciardi
Dr. Gabriel Molas

Author:
Flaminia Vinci
s300003

31/10/2024

Abstract

In the era of advanced technology , non-volatile (NV) *Resistive Random Access Memory* (ReRAM or RRAM) represents a revolutionary paradigm, able to withstand the Big Data requirement. Thanks to its intrinsic low power consumption, low cost, fast access times and high memory density, RRAM is forecasted to be a promising heir of flash technology.

Based on a simple two terminal metal-insulator-metal (in this work, said to be $TiN/HfO_2/Ti$) structure, its conductivity can be controlled by the application of a proper bias voltage during the programming phase, where the device moves from SET (HRS → LRS) to RESET (LRS → HRS) state.

This study revolves around the assessment of *Oxide-based RRAM* (OxRAM) retention capability at high (up to +150°C) operating temperatures. This is accomplished through Kinetic Monte Carlo (KMC) modeling, which takes into account the stochastic nature of resistive switching (RS). Conductive filament (CF) morphology variations and spread LRS resistance distributions are analysed and the long-term device performance is examined simulating *accelerated life tests* (ALTs). By focusing previously on the Ti/HfO_2 interface effects and, then, to CF dissolution due to V_O spreading, several annealing simulations are carried out for distinct baking times, up to 10 hours, and temperatures, up to 230°C.

Their effect, in conjunction with a right choice of activation energy parameters, is studied in terms of resistance drift distributions and QQ plots. Ultimately, the OxRAM lifetime is extracted through Weibull and Time-to-Failure graphs, exhibiting a slope of $E_A = 1.56eV$ and a maximum 10 years temperature of 106°C, which makes it a reliable device for typical commercial NVM applications.

To all those who
do not feel enough

Acknowledgements

To begin, I would like to express my sincere gratitude to Professor Carlo Ricciardi, Dr. Gabriel Molas and Dr. Boubacar Traore for their invaluable support and guidance throughout my six-month internship in Grenoble. Thank you for all the clarifications and advice that you gave me during my first professional experience. In this regard, I would like to thank also Professor Damien Deleruyelle, Professor Quentin Rafhay, Dr. Lucas Reganaz, Dr. Niccolò Castellani and all the colleagues for their collaboration and shared knowledge, who made this experience even more enriching.

This experience and my entire academic journey would not be possible without the support provided by my parents, who teach me the power of perseverance. Thank you to my sisters, Giordana and Domiziana, although distant, you are a safe haven and my point of reference. Thank you for always being by my side.
To my grandmother, my very first fan, who is able to find the right words, encourages me to believe in myself even more than I do. To my grandfathers, who watch over me from above like guardian angels. To my entire family, a certainty every time I come back home.

I would like to continue by thanking my best friend and "podcast sister" Silvia whose constant presence has been fundamental to my life. Thank you for all the adventures and silly moments we had together. Thank to my lifelong friend Martina, always ready to help and sustain with a nice word.
How could I not mention my dear friends and best roommates Claudia and Giusi. My life in Turin has been magical and it is all yours. Thank you for the laughter, the late-night talks, and the crazy moments, I will miss you very much.
I want to extend my thanks to my unique and irreplaceable colleague Giorgia. Despite the stressful moments and the chaos of projects, your support and friendship have made my academic life much lighter.
I would like to thank my squad: Francesco, Michele, Alex, Carmelo and even Luca; my experience at university would not have been the same without you. I will cherish the moments of joy and carefree in my heart forever. I want to thank the first people who welcomed me into our crazy life in Turin: Dario, Andrea, Luca and Riccardo. We will export our "Arancini receipt" everywhere.
I would like to thank my office colleague Khanh. Even though our time working together was brief, your presence made the work environment much more enjoyable, and I truly appreciate it. Finally, I would like to thank my adventure companions in Grenoble: Andrea and Vincenzo. You have been a wonderful discovery, thank you for all the coffee breaks and the deep reflections and meaningful conversations.

I can not conclude these acknowledgments without mentioning my greatest love, Gabriele. Thank you for the support you gave me every day through difficult times, even from hundreds km of distance. You are home no matter where we are. I feel incredibly fortunate to have met you, and this thesis is the result of the love you transmit to me every day.

Contents

1 Introduction to Emerging Memories **9**
1.1 Data growth . . . 9
1.1.1 Memory wall . . . 10
1.2 Non volatile memories . . . 10
1.2.1 Flash memory . . . 11
1.3 NVM - Emerging . . . 12
1.3.1 FeRAM . . . 13
1.3.2 MRAM . . . 14
1.3.3 PCRAM . . . 16
1.3.4 ReRAM . . . 16
1.4 In memory and neuromorphic computing. Why? . . . 18

2 Resistive Random Access Memory **19**
2.1 Resistive Switching . . . 19
2.1.1 The memristor . . . 19
2.1.2 Memory Resistive Switching . . . 20
2.2 ReRAM technology and integration with Logic . . . 21
2.3 ReRAM applications . . . 22
2.4 Oxide based ReRAM . . . 23
2.4.1 Redox reactions . . . 24
2.4.2 Oxygen ion and vacancy role in the ON-OFF switching mechanism . . . 25
2.4.3 Vacancy-assisted CF formation . . . 27

3 Kinetic Monte Carlo framework **30**
3.1 ReRAM modeling overview . . . 30
3.1.1 Microscopic models . . . 32
3.1.2 Device TCAD models . . . 32
3.1.3 Semi-empirical models . . . 34
3.2 KMC framework . . . 34
3.2.1 Parameters Initialization . . . 35
3.2.2 Generation, Recombination and Migration . . . 39
3.2.3 Monte Carlo event selection . . . 40
3.2.4 Electric Field and Electric Potential Poisson Solver . . . 41
3.2.5 Current calculation and conduction mechanism . . . 42
3.2.6 Compliance and Voltage drop . . . 46
3.2.7 Temperature update through Fourier Heat equation . . . 47
3.2.8 System Update and Stop of the Simulation . . . 48

4 Retention simulations **49**
4.1 What is retention and how it is calculated . . . 49
4.1.1 Aim of the simulations . . . 50
4.1.2 Bit error rate (BER) and failing bits . . . 51
4.2 Initial Setup: 100 OxRAM LRS programmed . . . 52
4.2.1 Activation energies role . . . 54
4.3 Simulation framework . . . 55
4.4 Ion-Interface Role . . . 55
4.4.1 Effect of Temperature . . . 55
4.4.2 Effect of Baking Time . . . 58
4.4.3 Ion diffusion within the resistive layer . . . 60
4.4.4 Ion diffusion within the scavenging layer . . . 61
4.4.5 Effect of absorption and emission barriers . . . 63
4.4.6 Failure probability . . . 64
4.4.7 Weibull graph . . . 66
4.4.8 Time to Failure . . . 67
4.5 Vacancy diffusion . . . 70
4.6 Two-step retention simulation . . . 71
4.6.1 First step: 1 minute relaxation . . . 71
4.6.2 Second step: Retention . . . 72
4.6.3 Weibull and TTF . . . 77
4.7 Single-cell LRS-retention with 3 V_O charge states . . . 79

5 Reset Programming and HRS retention **82**
5.1 Reset Programming . . . 82
5.1.1 Parameters Setup . . . 82
5.1.2 Reset: with and without recombination . . . 83
5.2 HRS Retention . . . 86

List of acronyms and abbreviations

1S1R One-Selector One-Resistor
1T1C One-Transistor One-Capacitor
1T1R One-Transistor One-Resistor
AI Artificial Intelligence
ALD Atomic Layer Deposition
ALT Accelerated Life Test
BE Bottom Electrode
BEOL Back-End of Line
BER Bit Error Rate
CB Conduction Band
CBA Cross-Bar Array
CBRAM Conducting Bridge Random Access Memory
CDF Cumulative Distribution Function
CF Conductive Filament
CG Control Gate
C2C Cycle-to-Cycle
CMOS Complementary Metal-Oxide Semiconductor
CNT Carbon Nanotube
D2D Device-to-Device
DBC Dirichlet Boundary Condition
DNN Deep Neural Network
DRAM Dynamic Random Access Memory
eRRAM Embedded Resistive Random Access Memory
FD Finite Difference
FDSOI Fully-Depleted Silicon on Insulator
FEM Finite Element Method
FeRAM Ferroelectric Random Access Memory
FG Floating Gate
FP Frenkel-Pair
HPQPC Hybrid Poisson Quantum Point Contact
HRS High Resistive State
IC Integrated Circuit
IMC In-Memory Computing
IoT Internet of Things
IPD Inter-poly dielectric
IT Information Technology
KMC Kinetic Monte Carlo
LRS Low Resistive State
MEM-RS Memory Resistive Switching
MEMS Micro Electro Mechanical System
MIM Metal-Insulator-Metal
ML Machine Learning
MLC Multi-Level Cell
MRAM Magnetoresistive Random Access Memory
MTJ Magneto Tunnel Junction
NBC Neumann Boundary Condition
NVM Non-Volatile Memory
OxRAM Oxide-based Random Access Memory
PCM Phase Change Memory
PDF Probability Distribution Function
PIM Processing in Memory
PZT Lead Zirconate Titanate
QQ Quantile-Quantile
QPC Quantum Point Contact
R/W Reading/Writing
ReRAM, RRAM Resistive Random Access Memory
RF Radio Frequency
RS Resistive Switching
SRAM Static Random Access Memory
SSD Solid State Driver
STT-RAM Spin Transfer-Torque Random Access Memory
TAT Trap Assisted Tunneling
TE Top Electrode
TMR Tunnel Magneto Resistance
TTF Time-to-Failure
VB Valence Band
VCM Valence Change Mechanism
VM Volatile Memory
W/E Writing/Erasing

Work environment

While nearing the end of my Master's program in *Nanotechnologies for ICTs*[1] at *Politecnico di Torino*, I had the opportunity to acquire hands-on experience through a six-month international internship at *CEA-LETI* in Grenoble, France.
During this period, I worked closely with *Weebit Nano*, a company that has rapidly emerged as one of the major players of the advanced semiconductor memory technology.
Founded in Israel in 2015, with the aim of building a revolutionary memory paradigm, in terms of storage and computing performances, *Weebit Nano* specializes in the development of state-of-the-art ReRAM (*Resistive Random Access Memory*) technology, an emerging type of NVM forecasted to be a promising heir of flash technology.
This is accomplished through the synergy of four main research division:

- Device physics;
- Process and materials;
- Analog design;
- Digital design.

Weebit Nano demonstrates its commitment to environmental sustainability by not using rare earth materials. Rather, the company focuses on utilizing fab-friendly materials that are easier to integrate into existing manufacturing processes.
The company collaborates with various industry partners and R& D institutions, among which CEA-LETI in France, where I was based.

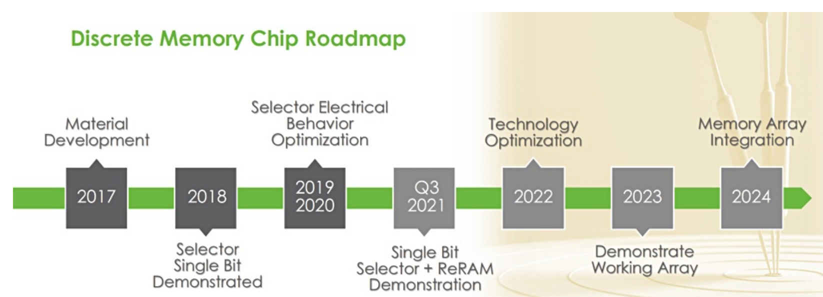


Figure 1: Discrete Memory Chip Roadmap, Weebit Nano. Adapted from [**1**]

The French *Alternative Energies and Atomic Energy Commission* (*CEA*) is a leading public research organisation, with nine headquarters, specialized in six main areas:

- Defence ad security;

[1]Information and Communication Technologies

- Low carbon energies;
- Technological research for industry;
- Matter and universe;
- Health and lifescience;
- Climate and environment.

Established in 1967, *CEA-LETI* is a department of CEA, belonging to the *Technological research division* (DRT), and working primarily on micro/nanotechnologies. It is globally accredited for bridging the gap between fundamental research and industrial applications, making it an ideal environment for the pursuing of pioneering technologies, such as those brought by Weebit Nano.
My experience greatly benefited of this highly collaborative atmosphere between Weebit engineers and CEA researchers, all coming from diverse disciplines, but working closely with the common objective of driving innovation in ReRAM technology.

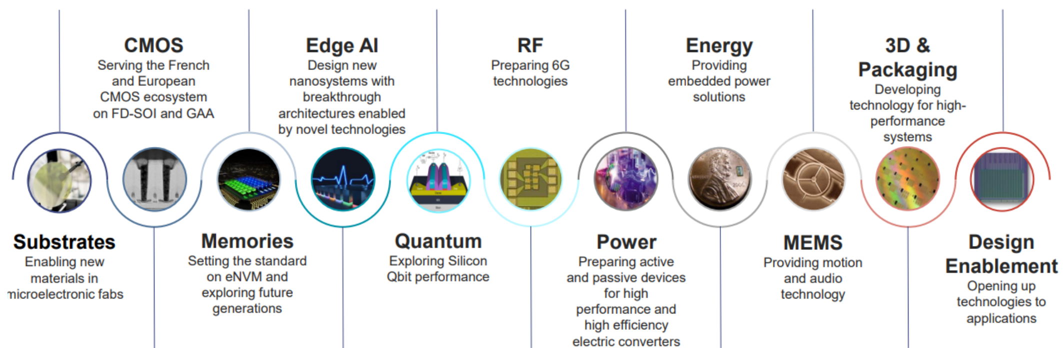


Figure 2: Research areas of DCOS (Silicon Composites) department, CEA-LETI

Introduction to Emerging Memories

This chapter seeks to introduce the reader to the emerging memory technology world, emphasizing its role in proposing revolutionary memory paradigms. It discusses how the Von Neumann bottleneck has been overcame, along with a much lower energy demand.
The advantages and disadvantages of this technology will be reviewed and a comparison with today well-established Flash memory is discussed.
Finally, the employment of ReRAM for in-memory computing and neuromorphic applications will be mentioned.

1.1 Data growth

In the era of advanced technology, the emphasis is on examining what is new in the market and the potential updates it brings, enabling the user to capitalize on them.
This elucidates, for instance, the significant research made in AI, machine learning and IoT, in general, as companies strive to connect devices and gather data for enhanced functionalities and insights.
On the other hand, having more and more devices connected all around implies an exponential growth of their data, up to more than 180 zettabytes in 2025 [2]. As a consequence, this poses some limitations on the data processing infrastructures, either software or hardware.
In particular, ultra-low power and "always-on" devices are required for instant data generation. Whereas, Big Data spreading demands for a large communication bandwidth, together with computing and memory resources to be able to catch-up with its data-intensive applications.

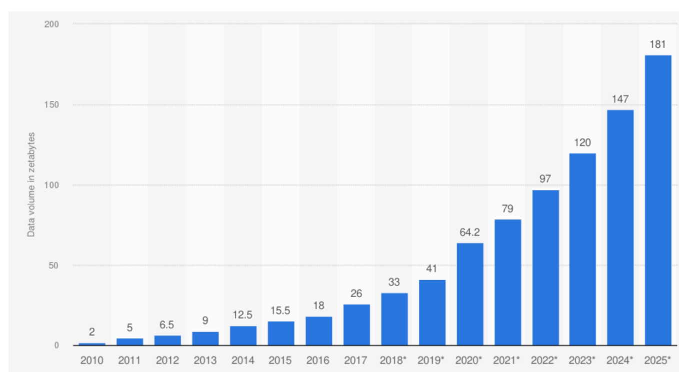


Figure 3: Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025. Adapted from [**5**]

1.1.1 Memory wall

To accomplish the huge data call, from the hardware standpoint, the standard is to exploit multi-core architecture for processors, pursuing on with scaling.
At the same time, the rate of improvement in DRAM memory speed lags significantly - with an average decrease of 30% every two years[8] - behind the processor core's speed, leading to what is known as the *memory wall* or *Von Neumann bottleneck*.

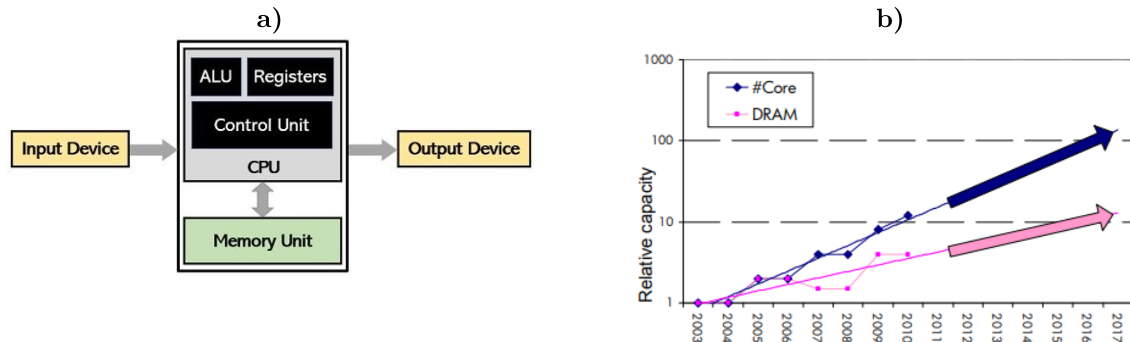


Figure 4: a) Von Neumann Architecture, adapted from **[11]** b) Year trends of memory and processor capacity. Adapted from **[8]**

The origin of this performance gap comes from the fact that, according to the Von Neumann Architecture, modern computers are organized into two separated components: a processing unit (CPU) and a memory unit. If, on one side, the number of transistors integrated in a computer chip doubles every 18 months, in line with Moore's law, on the other side, the same cannot be said for the memory module.
As a consequence, data transfer between the two units, ensured through a BUS, is limited in terms of bandwidth. Furthermore, latency effects arise due to the fact that data must be physically moved from one element to the other, and vice versa, every time that a reading or writing operation is performed. This issue can be partially mitigated through the exploitation of the memory hierarchy principle. In this context, small cache memories, typically SRAMs, are directly integrated inside the processing unit, for hosting frequently accessed data.
However, the scaling problem persists, since cache memory sizes are not able to keep up with transistor scaling, otherwise the overall yield would be affected.
In addition, the larger separated memory component is usually based on DRAM, and the aim, today, is to push forward with footprint scaling, down to the practical limit of 4F2 [6].
Not only do volatile memories (VM) suffer from the bottleneck, but also nonvolatile (NVM) flash memory, exploiting CMOS structures for temporary and permanent data storage, are limited in terms of size and density.
The principle and physical mechanism of Flash memory will be further investigated in Section 1.2.1.

1.2 Non volatile memories

Throughout the thesis, a comprehensive emphasis is posed on non volatile memories, characterized by the ability of retaining data even when the power supply is off.
NVM can be classified into two macro categories:

- Charge based memories (e.g. Flash);
- Non-charge-based-storage memories, also known as *Emerging Memories*.

Nowadays, the 99% of applications, ranging from consumer electronics to Solid State Drives (SSDs) and USB flash drives, is made of Flash memories [6].
Nonetheless, emerging memory technology is making headway into the market and is expected to be a key enabler for big data systems, meeting also the temperature requirement of the automotive field.

1.2.1 Flash memory

Flash technology is considered as the cornerstone of non-volatile memory. Thanks to its high maturity, flexibility, and overall performances, it is employed as benchmark for the study of other, emerging, NVMs.
Its simplest structure is the NOR flash type, developed first in the 80's just for read-only purposes, and characterized by fast random access times, at the expense of slower erase and write cycles[13].
Later on, the NAND counterpart was launched into the market, and, thanks to its larger density and storage capacity, it was adopted as the main option for applications involving large amounts of data[13].

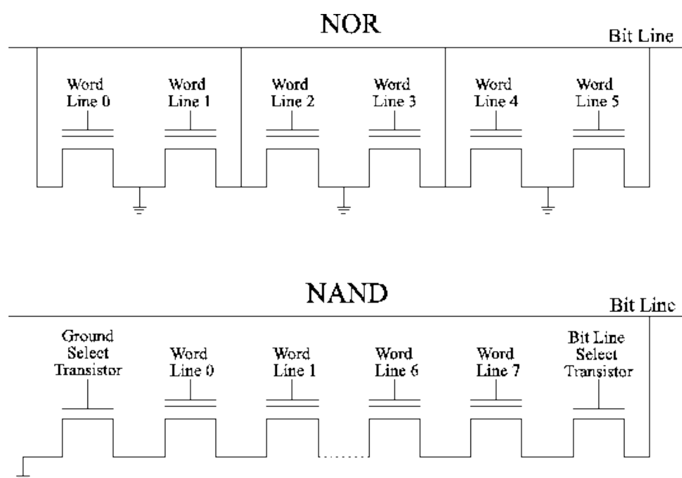


Figure 5: Basic circuits for NAND and NOR flash memory devices. Adapted from **[13]**

Single memory cells are arranged into an array, and each of them include one floating gate transistor (1T) which serves both as selection and storage node. This technology is the dominant one for 2D flash memories,and is still used for 3D memories by Intel and Micron[67].
By controlling properly the threshold voltage V_{TH} at the floating gate, Multi-Level Cells (MLCs) can be realized, meaning that more than 1 bit of information can be stored in a single cell.

Cell type	Bit stored	Levels of info	Endurance
SLC	1 bit	2 levels	$5 \cdot 10^4 - 10^5$
MLC	2 bit	4 levels	$5 \cdot 10^3$
TLC	3 bit	8 levels	$3 \cdot 10^3$
QLC	4 bit	16 levels	$< 10^3$

Table 1: Different types of multi level cells and their respective endurance. Adapted from **[14]**

Due to physical and technological constraints, having multi-bit cells reduce the voltage margin between levels, making them more sensitive to unwanted trapped charges.

Working principle

Referring to Fig. 6, in the gate area, from top to bottom, four main layers can be distinguished[15]:

- Control gate (CG), usually made in PolySilicon;
- Inter-poly dielectric (IPD), consisting of a stack of $SiO_2 - Si_3N_4 - SiO_2$ (ONO);
- Floating gate (FG), made in polysilicon;
- Tunnel oxide (SiO_2).

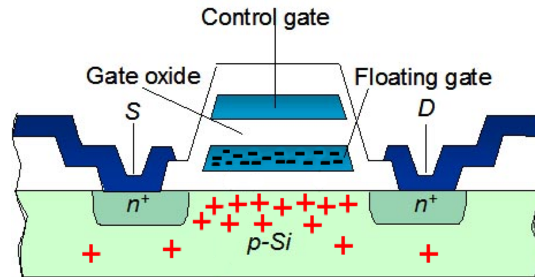


Figure 6: Basic scheme of a flash memory cell. Adapted from **[16]**

Here, programming is accomplished via charge trapping. Indeed, by applying the right voltage to the CG, carriers may tunnel from the channel, through the oxide layer, to the FG, and vice versa for the erase operation. The charge state of the floating gate (FG) is, then, translated into a source-drain current I_{SD}, that is used as a criterium for determining the presence or absence of the information bit.
Since the FG is electrically isolated by top and bottom oxides, even when the power is off, charge remains trapped inside, and the non-volatility requirement is satisfied.

Challenges and limitations

Regrettably, it can occur that a non-negligible number of carriers tunnel from the FG into the inter-poly dielectric (IPD), thereby diminishing the overall endurance[1] of the flash memory (10^4 cycles vs 10^{15} cycles of SRAM and DRAM).
The leaking effect is further enhanced with device miniaturization, since the FG, IPD and oxide thinning leads to a decrease of the energy barrier amplitude for overcoming the three different layers [16].
This scaling limitation prevents FG-MOS downscaling beyond the sub-10nm technology node, otherwise the threshold voltage control is not anymore ensured, together with a too limited number of charge trapped in the FG[15].

1.3 NVM - Emerging

Although advancements have been made with 3D vertical Flash and there is a trend towards using low-dimensional emerging materials like CNTs [15], the consumer electronic market is shifting toward more "data-centric" applications [2], for the reasons outlined hitherto.

[1]endurance is the number of times a block of memory can be programmed and erased (P/E cycles) before the memory wears out and stored data becomes unreliable.

[2]replacing high performance CPU-centric architectures

Therefore, a non-charge based alternative is needed, that satisfies the requirements of the beyond-CMOS era, such as low power consumption, low cost, high memory density and novel functionalities. In this context, NV emerging memories are the perfect candidate for big data systems, thanks to their byte-addressability, small access times and the possibility of building high-density cross-bar arrays (CBAs).
The name *emerging* comes from the fact that they are least mature technologies and, still today, they are significant object of study for the R&D community.
As a matter of fact, some of these memories have already reached the production line[5], and they pertain to the *prototypical* category of NVMs. This includes ferroelectric random access memory (FeRAM), magnetoresistive RAM (MRAM), phase-change memory (PCM), spin transfer-torque RAM (STT-RAM), and resistive RAM (ReRAM).
In the next sections, the peculiarities of each type will be briefly examined.

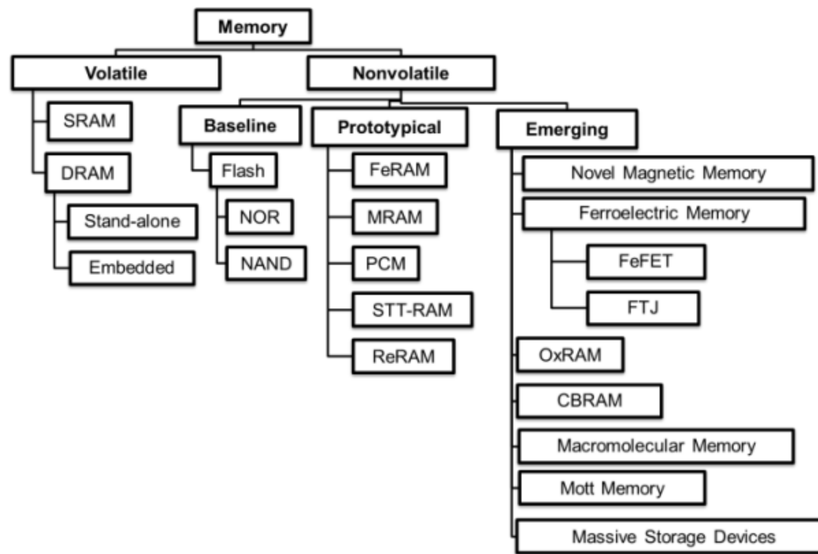


Figure 7: Taxonomy of memory devices. Adapted from **[5]**

1.3.1 FeRAM

The FeRAM cell is constituted by a 1T-1C (one-transistor one-capacitor) structure, as in DRAM, with the difference that, for the FeRAM, the capacitor dielectric layer is replaced by a ferroelectric material, such as pervoskites (PZT[3]) or transition metal oxides like HfO_2, which is directly integrated inside the BEOL of 130nm CMOS process[18][68].
Ferroelectricity is the ability to get a spontaneous polarization when subjected to an external electric field. By properly controlling this native polarization, the memory cell is able to process several information levels.
The electric dipoles, trapped inside the thin crystalline film, change polarity according to the sign of the applied voltage, thereby inheriting a binary switch.

[3]lead zirconate titanate $PbZr_xTi_{1-x}O_3$

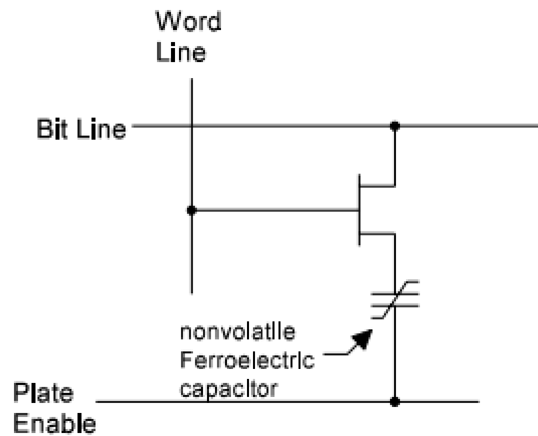


Figure 8: 1T1C FeRAM cell. Adapted from **[17]**

FeRAM technology draws attention in the NVM market, outperforming the conventional Flash one, because it brings various innovations:

- Energy efficiency, since low programming voltages are enough to displace charges;
- Fast read/write access time, occurring within a few hundreds of nanoseconds, without affecting data integrity;
- Large endurance levels, of the order of 10^{13} W/E cycles (*Fujitsu*[18]);
- Robustness against disruption and magnetic interference, further enhancing FeRAM reliability and lifetime, particularly in harsh environments, such as space.

Nevertheless, FeRAMs are characterized by a read-destructive operation. Indeed, data just read are inherently erased, recalling the need of writing them back to prevent permanent loss or corruption.

1.3.2 MRAM

Unlike semiconductor memories that leverage electronic charge to store data, for MRAMs, the electron spin is determinant for storing information. The primary benefit of utilizing this technology is that magnetic domains are inherently stable, as opposed to charge, thereby eliminating the requirement for constant power delivery.
MRAM has the potential to be addressed as the "universal memory" [6] because of its capability to combine the speed of SRAM and the density of DRAM, together with the non-volatility of flash memory.
Other interesting advantages are:

- Small R/W access times (down to few ns [19]);
- Unlimited endurance, up to 10^{16}-10^{18} write cycles [20];
- Long-term data retention, > 20 years at 125°C [20];
- Possibility of arrange 3D structures.

On the contrary, disadvantages are outlined:

- Scale limitation beyond 65nm technology node[6] for the MTJ MRAM, because of a too high magnetic field, which could lead to electromigration issues;

- Currently commercially available STT-MRAMs are still not as fast as DRAMs, making them suitable only for embedded applications;
- Complex technological structure.

Regarding the last point, in principle, the core structure is a simple device, since it is made of three main layers, each of them constituted of several sub-films. The issue, however, is that these layers are nanometer-thick and a high surface quality is required, making deposition and etching not so trivial. At the same time, patterning lithography is not a concern since, in addition to each cell, a transistor, typically a FDSOI, is put in series, making, as said, the area larger.
The basic magneto-tunnel junction (MTJ) stuck is made of two ferromagnetic layers, one with fixed and the other with free magnetization, with an oxide (usually a transition metal oxide like AlO_x before and MgO_x then) in the middle, acting as insulating tunnel barrier.

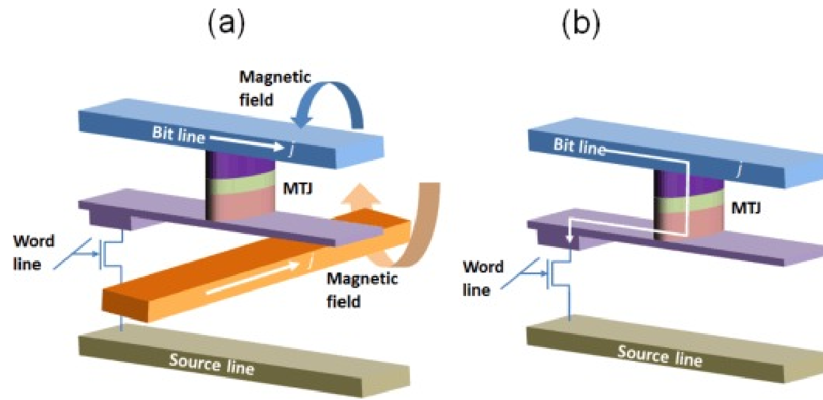


Figure 9: Schematic illustration of (a) field-driven switching MRAM and (b) STT-MRAM. Adapted from [**18**]

Through the exploitation of the tunnel magneto resistance (TMR) effect, the value of resistance is determined by the magnetic moment orientation of the free layer with respect to the fixed layer.
Simplifying[69]:

$$TMR = \frac{R_{MAX} - R_{min}}{R_{min}} = \frac{R_{AP} - R_P}{R_P} \tag{1.1}$$

where R_{AP} and R_P are, respectively, antiparallel and parallel resistance.
In the magneto tunnel junction (MTJ) memory element, a tunneling current is generated, as a result of the induction of two orthogonal magnetic fields at the electrodes (respectively named bit line and word line).

STT-MRAM

Next generation spin-transfer torque (STT) MRAM have the potential to reduce cell footprint and power consumption (Fig. 9(b)).
The H-field is replaced by a spin-polarized current able to transfer its angular momentum to the free layer. This can significantly reduce MRAM write currents and self-heating.
Perpendicular magnetization[4] and scalability below 32nm, with STT approach, are possible, thanks to a simpler technology structure[70].

[4]out-of-plane anisotropy

1.3.3 PCRAM

Phase change memories belong to the "resistive device" category, due to the significant change in resistance used to select the bit of information. To be specific, this resistance variation is the result of a transition between the amorphous and the crystalline phase of a *phase change* material, owing to an external stimuli, typically heat, in the form of an applied voltage or current at the electrodes.
Typical PCMs are chalcogenide-based glass like GST[5] alloys, properly doped with nitrogen and/or oxygen impurities[21].

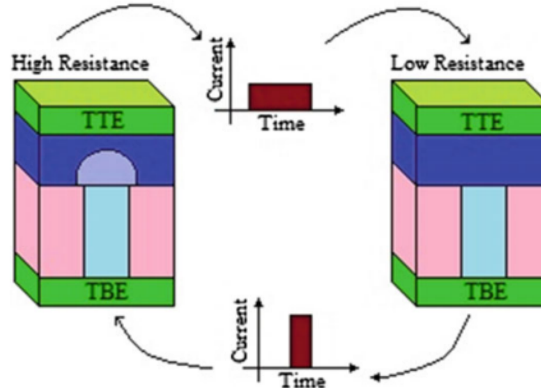


Figure 10: Schematics of resistance switching in PCRAM. Adapted from **[21]**

As illustrated in the figure, two states can be underlined in the context of PCRAM.

- HRS (*High resistive state*): corresponding to the amorphous phase of the material, it is reachable by applying a high temperature (600°C) for few nanoseconds, to melt and solidify fastly the PCM[22]. In this case, the cell is said to be in *reset* state.
- LRS (*Low resistive state*): corresponding to the crystalline phase, it is achievable by applying a relatively low T, but still higher than the crystallization one T_C, for long time. In such circumstances, the cell moves to the *set* state.

Together with that, discrete intermediary states can be attained, allowing for the realization of multi-bit cells.
Finally, in order to read the stored information, a small voltage pulse is applied to the circuitry.
In this regard, it is worth noting that, since the physical structure of the device undergoes a large stress during phase transitions, PCRAMs suffer of low endurance level (around 10^8 W cycles for the standalone device [23]).
For what regards the speed, this is dictated by the set phase. The shorter is the pulse width, the faster will be the writing speed.
The reset state, instead, is crucial for the final power consumption, since the single cell can be highly power demanding. This, in turns, affects the memory density, because a selector, able to deliver such high currents, is adopted.

1.3.4 ReRAM

The technology that has revived the interest of the current study is Resistive RAM, abbreviated ReRAM or RRAM, which falls into the category of memristors.

[5] $Ge_2Sb_2Te_5$

Here, the bit of information is determined by the value of resistivity of a simple two terminal metal-insulator-metal (MIM) structure, similar to a parallel plate capacitor, which makes such technology so attractive, especially for stacking 3D arrays.
In order to make the device ready for data storage, a "soft breakdown" is induced as initial stage, called *Forming*. The latter is accomplished by applying a relatively large voltage at the electrodes, so that a conductive path between the two is formed.
After the configuration step, the resistance of the pristine cell can be controlled through the application of proper currents and voltages. This is the programming phase, where the device moves from SET (HRS → LRS) to RESET (LRS → HRS) state, encoding respectively bit '1' and bit '0'.
In accordance to this, the conductive filament (CF) formed inside the switching (insulator) layer, repeatedly forms and dissolves, thereby permitting or obstructing the flow of current.

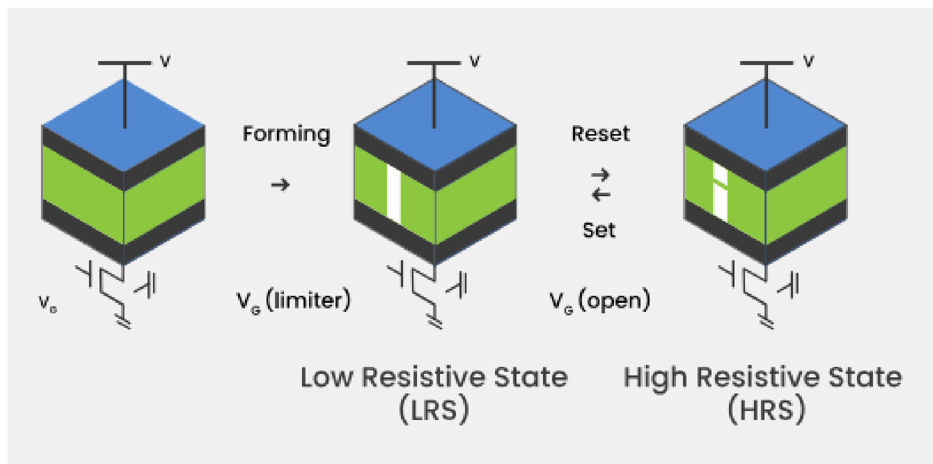


Figure 11: Schematic illustration of a ReRAM cell and its configuration during pristine, set and reset state. Adapted from **[24]**

Behind this programming/erasing cycling, there is a persistent movement of ions that ultimately degrades the cell over time, rendering it not reliable anymore. Indeed, endurance and retention are very crucial, and a trade-off between the two is required to make the device ready for commercial distribution[35]. In this scenario, the right material choice plays a fundamental role.
Several ReRAM options are actually possible, and can be categorized into two large families:

- CBRAM, which stands for *Conducting Bridge RAM*, exploiting the electrochemical control of metallic inclusions migrating inside a thin electrolyte layer, sandwhiched between two electrodes, one inert (e.g. W, Ta, Pt) and one made of an electrochemically active material (e.g. Ag, Cu, Ni)[5].
  CBRAM-based products are, today, commercially available, with the first 16 GB $CuTe$ CBRAM array being introduced in 2014.[5]

- OxRAM, standing for *Oxide-based RAM* and combining the oxygen ions migration together with redox process at the electrode interface and resistive layer, usually made of transition metal oxides like TaO_x and HfO_x.

In addition to these, there is a third category, which is however still at the research level, known as Macromolecular (polymer) Memory. Here, as the name suggests, the layer interposed between the metal electrodes is a polymer containing nano-particles, whose transfer is decisive for determining the resistive state.
In Chapter 2 the oxide-based resistive memory (OxRAM) is investigated in detail.

1.4 In memory and neuromorphic computing. Why?

The reason why the emerging memory approach has been discussed so far, is to display it as a viable solution for mitigating the data transfer latency between the CPU and the external memory.
Placing the storage element closer to the processing unit is not enough and, as a matter of fact, the concept of *In Memory Computing* was born in the early 2010s. Since that time, advancements beyond the basic caching organization have been made, and, currently, the aim is to relocate the main data storage element, namely the DRAM, within the processing unit. *Processing in Memory* (PIM) not only reduces latency, but also favours data-intensive applications and real-time analytics.
This is where emerging memories comes into the picture, and also thanks to the progresses done in AI/ML, PIM-based devices are becoming a reality, today.
The next evolution of this idea is to replicate *in-silico* how the human brain works and, most importantly, its great energy efficiency, with a dissipation of just 20W [25], in contrast with the CMOS-based Von Neumann Architecture, forecasted to reach 10^{27} Joule in 2040. [26]
Among the different emerging NVMs outlined up to now, embedded ReRAM seems to be the promising candidate for the creation of a new architecture able to co-locate logic and memory within the same area and to exploit that for *neuromorphic applications*.
Thanks to its very basic structure, ReRAMs show an acceptable price competitiveness and offer a large integration density, making it suitable for big-data management. Further remarkable points are:

- Low power consumption;
- Fast access times for reading and writing (≈300 ps[43]);
- Good endurance;
- Easy to be integrated into the CMOS process workflow, further contributing to reduce latency.

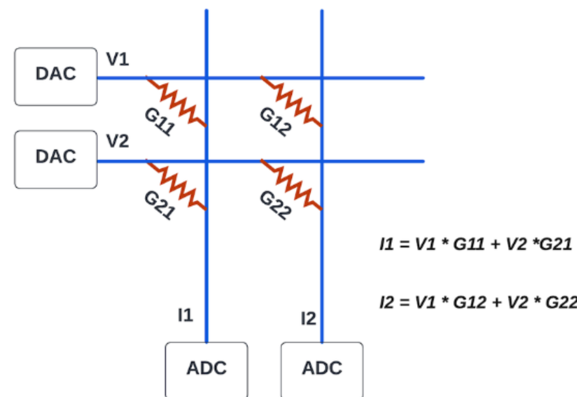


Figure 12: Basic scheme of a RRAM crossbar array, employed for MVM in DNNs. Adapted from **[27]**

Taking as an example Deep Neural Networks (DNNs), their working principle relies on the convolution between multiple inputs from the previous layer and the weights from the current layer. In conventional GPUs these weights occupy the primary memory and are fetched for each layer's computation.
Conversely, by looking at the memory architecture as a crossbar array of resistive elements, each of them able to store the weight, due to the analog nature, the achievable density is much larger.
Nevertheless, the employment of ReRAM devices for IMC and neuromorphic applications goes beyond the scope of this work. It is mentioned solely to inform the reader about the significant efforts being made by the IT industry in this direction.

Resistive Random Access Memory

ReRAM is believed to be the leading candidate to replace flash memory in the next future. Its history, technology, and large field of applications are further investigated in the following Chapter.
The focus of this work is on the OxRAM type, whose physical mechanism is examined through ab initio insets.

2.1 Resistive Switching

2.1.1 The memristor

The history of memresistive devices, intended as two-terminal resistive switching components, dates back a long time.
The first to introduce this new concept was Professor L. Chua in 1971 [28]. He was the pioneer of the *memristor* as fourth basic circuit element, further evolved under the more general term of *Resistive Switching* through the work of Professor R. Waser in 2007 [30].
The key distinction between a memristor and a traditional resistor lies in the IV characteristics. Unlike the linear ohmic behavior described by Ohm's law, the memristor exhibits a curve pinched at the origin. This is fundamental for memory purposes, because it means that, depending on the applied voltage, two stable states can be reached, either with high resistance (OFF, bit '0') or low resistance (ON, bit '1'), in a reversible way, maintaining trace of the state before.
The latter is the basic principle of synaptic plasticity and, for this reason, memresistive devices are exploited for machine learning applications.

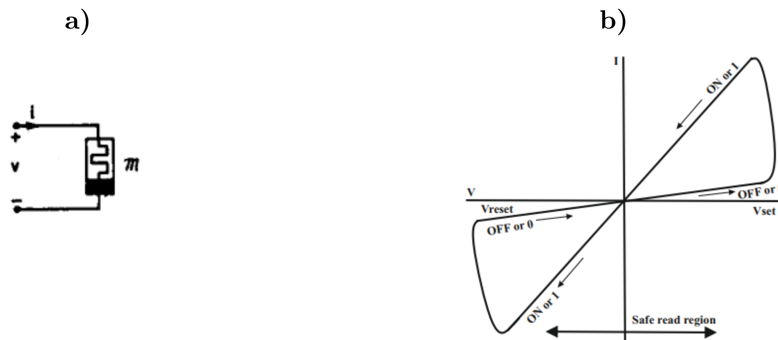


Figure 13: a) Proposed symbol for memristor by Prof. Chua, adapted from **[28]** b) Memristor I-V curve in linear scale, adapted from **[29]**

2.1.2 Memory Resistive Switching

Nanoscale resistive switching devices found applications either for logic and nonvolatile memory fields; the latter being the object of this study.
As stated, the memristor's feature highlights a dependency on a state variable, which is considered to be time, since the internal state is influenced by the current-voltage history. Such state is maintained over time, allowing to fulfill the nonvolatility requirement.
In his "*Nanoionics-based resistive switching memories*" [30], R. Waser defines Resistive Switching (RS) as any phenomenon which relies on the variation of resistance at the nanoscale level, by a change in the atom configuration the MIM stack is made of, to realize a binary, or multinary, switch. Such change at the microscopic level is possible by means of voltage pulses with controlled amplitude and width, and it can be of various nature, depending on the dominant contribution, either: thermal, electronic or ionic.
Memory Resistive switching (MEM-RS) comes into play when such RS stimulus affects a state variable somehow related to the resistance, whose value is memorized inside the element.
Taking as reference Fig.13b, two separate resistance levels can be observed:

- HRS: High Resistance State, or OFF state (low conducting), corresponding to bit '0' of information;
- LRS: Low Resistance State, or ON state (highly conductive), corresponding to bit '1'.

Furthermore, two types of switching polarity are possible:

- Unipolar, or symmetric, switching, where V_{SET} and V_{RESET} have the same polarity but different amplitude;
- Bipolar, or antisymmetric, switching (Fig.13b), where Set and Reset programming depend on voltage polarity. As the name suggests, in this case, from a material point of view, a sort of asymmetry must be introduced, playing on the stoichiometry of the insulating layer.

Ultimately, in both scenarios, reading can be accomplished with low magnitude voltages, ensuring the stability of the state.

Figure 14: a) Schematic of a single ReRAM cell with applied voltage. b) Organization of ReRAM cells into a crosspoint memory array ensuring high-density and low-power storage. Word lines and bit lines are utilized for selecting a memory cell as well as for writing and reading data. Adapted from **[31]**

2.2 ReRAM technology and integration with Logic

Before delving into the device's physics, its material composition and fabrication process will be discussed.
All RRAM typically show a 1S1R configuration, where the selector (S) is, in this case, a CMOS transistor (1T1R), whose gate is properly opened for reading a specific memory cell in the cross-point array.
RRAM has been selected as the promising candidate for replacing Flash memory, and the reason of that has to be found in the improvements reached in terms of density storage, fast access times, low power, good endurance and cost effectiveness.
Nevertheless, the IT research community continues to encounter multiple challenges, with a prominent issue being the inability to achieve high endurance and high retention times simultaneously.
In Ref.**[35]** it has been put in evidence how a trade-off between the maximum achievable number of R/W cycles which can be performed on a single memory cell and the maximum time a single state is maintained, is always needed.
Among the several choices, the material stack fulfilling this compromise is made of TiN/HfO_2/Ti, with a oxygen vacancy based filament. In detail, as reported by the scheme of Fig. 15, for an individual cell, endurance levels up to 10^8 cycles can be reached, along with retention times of 10^5s at a maximum temperature of 200°C[35].

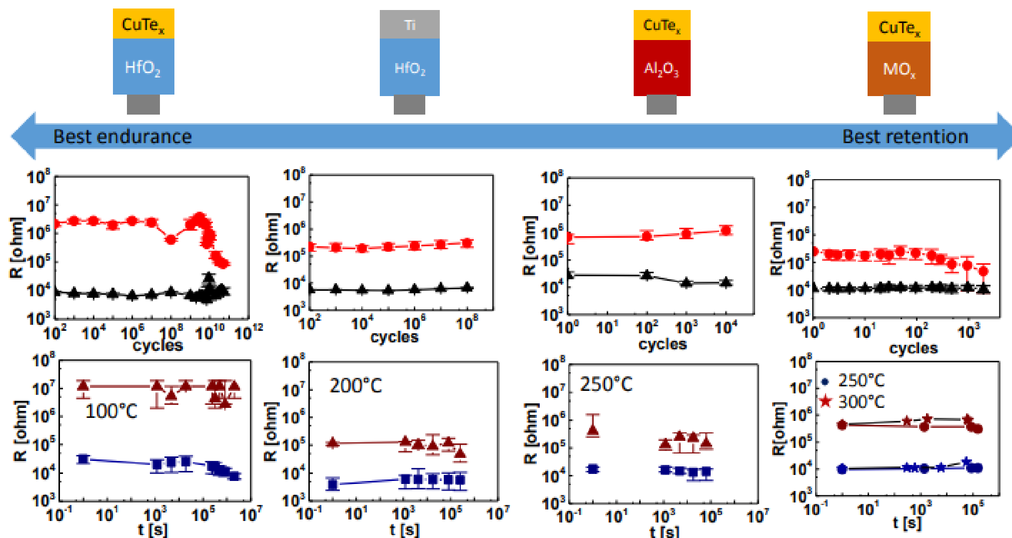


Figure 15: Endurance and retention behavior with a given window margin WM=10^2, for 4 different RRAM classes. Adapted from **[35]**

Another noteworthy point which makes embedded ReRAM technology so appealing among the future memory generations, is that, due to the very basic structure, it can be easily integrated in the back end of line (BEOL) 130nm-CMOS or 28nm-FDSOI process flow [36] above the logic area.
This makes its manufacturing very cost-effective, since the addition of 2 mask layers[32] is enough for making seamless high density integration possible (up to 10^{11} bits cm^{-2} [43]), exploiting the same metal and insulating layer of the BEOL. On the contrary, flash memory can be built only in the front end of line (FEOL) portion, with consequent increase of fabrication complexity.

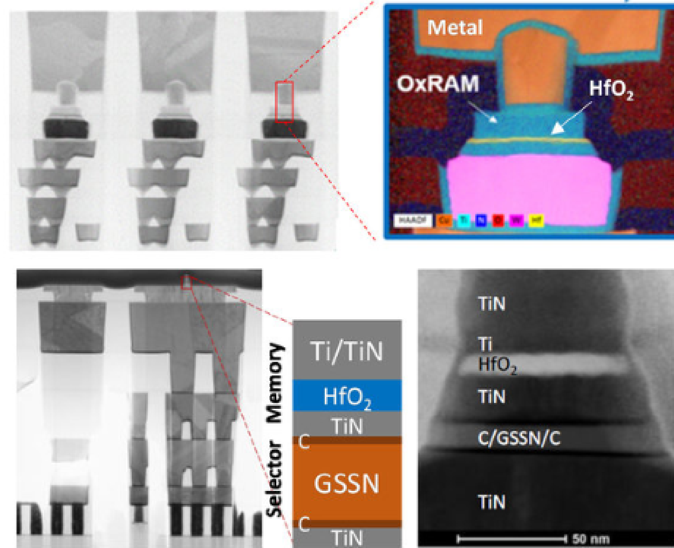


Figure 16: Transmission Electron Microscopy(TEM) cross sections of a HfO_2 based OxRAM integrated in the BEOL of a 28 nm FDSOI technology (top), and an OxRAM/OTS device integrated in the BEOL of a 130 nm CMOS process(bottom). Adapted from **[36]**

The reference structure for the study conducted is a 5nm ALD film of HfO_2, sandwhiched between a bottom TiN layer ad a top Ti/TiN electrode, deposited through PVD and with the scavenging TiN layer 5nm thick as well[34].
When integrated into a 130nm CMOS node, the final area is 300nm x 300nm , while, for the 28nm FDSOI choice, the area decreases to 160nm x 160nm[46].

2.3 ReRAM applications

Pursuant to the final application, NVMs can generally be categorized as either embedded or standalone memories.
For what regard the resistive memory, embedded RRAM (eRRAM) has experienced a greater commercial development with respect to the standalone counterpart, owing to various technological advantages influencing the market today. The roots explaining this trend have to be found in a growing demand for integrated highly-scalable solutions able to enhance the performance capabilities of SoCs architectures.
The following is a list of potential applications for which eRRAM can be suitably tailored[38][46].

- Due to its low power consumption and great scalability, eReRAM is a key factor in developing energy-efficient IoT systems. This is particularly relevant for mobile devices and wearable always-on technology, where battery life is crucial, also in harsh environmental conditions.
- When integrated within the same die as logic circuits, the distance that the signal has to travel is minimized, reducing, along with, latency and heat dissipation. Applications requiring efficient data processing capability, such as real-time data analytics and edge computing, benefit from this feature.
- The analog/mixed-signal nature of RRAM is leveraged in analog ICs, such as power management ICs (PMICs), display driver ICs, RF, MEMS, and others.

- The automotive market is investing a lot on embedded NVM, employed for the hundreds of smart sensors that, today, are present in the vehicle. This field of application is one of the most challenging still today, since reliability for temperatures ranging from -40°C up to 150°C[39] [1] must be ensured. However, good results of 10 years retention over 125°C have already been qualified (*Weebit Nano*[40]).
- eRRAM exhibits promising results even in highly radiative conditions, making it a suitable technology for aerospace and medical devices.
- The ability of ensuring fast switching between multi-level bits without affecting data integrity, makes eReRAM a valuable instrument for AI and ML applications, contributing to the neuromorphic research. Indeed, another noteworthy point is that the dynamics of the RS process in a ReRAM is governed by ions[2] flow rather than electrons, the redox reactions they participate in. This is important in the context of DNNs, as formation transfer between synapses is enabled by ion (Ca^{2+}) passage[71].

2.4 Oxide based ReRAM

As anticipated in Par.1.3.4, according to the nature of the ionic species, two classes of filamentary ReRAM can be discerned:

- CBRAM, mediated by cation dissolution and successive migration from an electrochemically active electrode;
- OxRAM, driven by anionic flow and filament creation upon V_O generation.

Let's take as reference cell the scheme of Fig.14a, and let's focus on the OxRAM type. When a positive voltage is applied to the anode, negatively charged oxygen ions (O^- or O_i) migrate towards the anode, leaving behind electron-deficient oxygen vacancies (V_O^+), drawn to the negatively charged cathode.
The switching process is mediated by V_O, which are responsible of the formation and disruption of conductive channels within the oxide, which is usually a transition metal oxide (TMO) like TiO_2 and HfO_2.
This phenomenon is referred to as the *Valence Change Mechanism* (VCM) because, when $O_i - V_O$ migration occurs, it results in an alteration of the oxide stoichiometry, in turns observed as a variation in the electronic conductivity at the electrode interface.

[1] Grade-0 of the automotive market
[2] from which the name "nanoionics-based resistive switching"

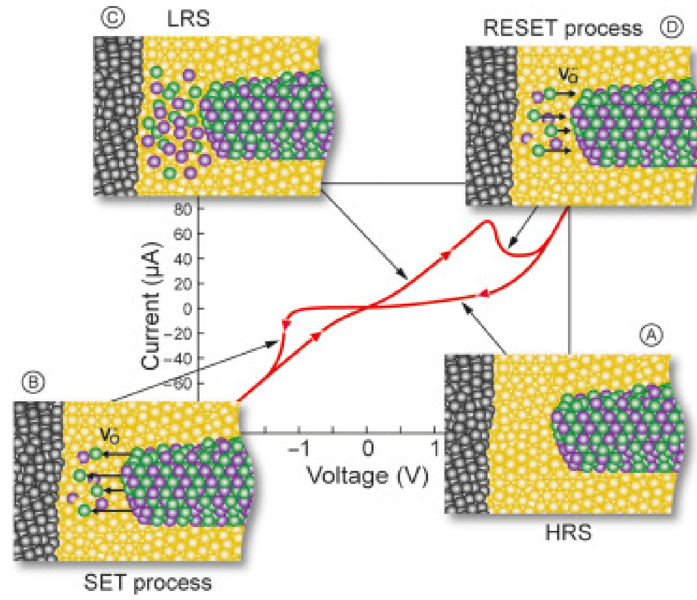


Figure 17: CF formation and disruption, with corresponding IV behavior, during SET and RESET operation in an OxRAM cell. In green are shown the oxygen vacancies, in purple and yellow are represented the Hf ions in low and standard valence state. Adapted from **[41]**

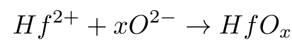
2.4.1 Redox reactions

In order to foresee the reliability characteristics of an OxRAM, it is essential to understand the link between the MIM physics and the device electrical behavior.
Among the numerous proposed theories, the most accreditated hypothesis involves redox reactions responsible for the generation and recombination of O_i/V_O pairs at both the insulator and the electrode/insulator interface level.
For what regard the RS medium, Hafnia in this work, the electrochemical redox reactions it is involved into, are resumed in the following.

- For the reduction of stoichiometric oxide:

$$HfO_2 \rightarrow Hf^{4+} + 2O^{2-}$$

- For the non-stoichiometric oxidation:



$$Hf^{2+} + xO^{2-} \rightarrow HfO_x$$

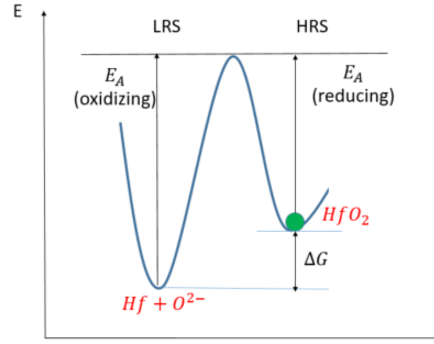
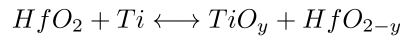


Figure 18: Sketch of the potential energy curve of HfO_x redox pair. The stability of one resistance state over the other is ensured by the stability of the redox pair.

Whereas, as confirmed by the experimental and first-principles calculations conducted in Ref.[**34**], due to the higher oxygen reactivity of the top Ti electrode, with respect to the more inert BE, made of TiN, redox reactions can take place also in correspondence of the TiN/HfO_2 interface, leading to a substoichiometric region 2-3nm thick, described by the reaction[34]:



$$HfO_2 + Ti \longleftrightarrow TiO_y + HfO_{2-y}$$

Thermodynamic enthalpy studies support the diffusion of oxygen from the insulator to Ti for a minimum energy barrier of 1.5eV, ending up with a most probable composition of TiO_2/Hf_2O_3.[34]

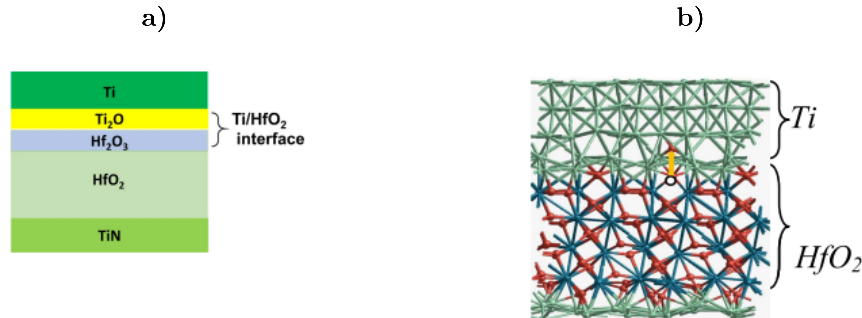


Figure 19: a) Proposed sketch of the most propable TiN/HfO_2/Ti stack, showing interface layer b) Ti/HfO_2 interface ab initio model. Adapted from [**34**]

2.4.2 Oxygen ion and vacancy role in the ON-OFF switching mechanism

Oxygen vacancies have been identified as the main actors ruling the OxRAM switching mechanism. These are defects in the oxide that arise when a specific voltage is applied at the metal contacts. Specifically, this type of defect is known as a Frenkel-Pair (FP) defect of anionic nature, as an oxygen ion occupies an interstitial position in the lattice, leaving behind an oxygen vacancy.
Consequently, this results in the creation of a I_O/V_O pair due to the breaking of $Hf - O$ bonds.

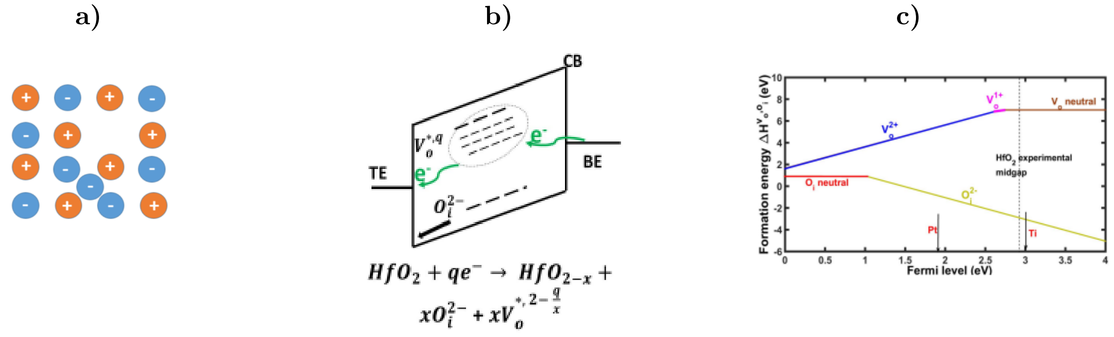


Figure 20: a) Anion Frenkel-Pair defect. b) Energy band diagram of the BE/HfO_2/TE upon electron injection from the BE. Adapted from **[34]** c) V_O and O_i formation energy in HfO_2 vs Fermi level, measured form VB. Adapted from **[34]**

Thermodynamic stability analyses from Fig.20c) indicate that oxygen ions are more energetically favorable over the O_2 molecular form.
Whereas, regarding vacancies, three types can be identified: doubly charged, singly charged, and neutral V_O, depending on the voltage extent. Their role is better discussed in Par.2.4.3.
Considering the band diagram depicted in Fig.20b) for a positive bias, an increase in this bias enhances electron injection from the BE and makes it more predisposed to subsequent tunneling at V_O energy levels below the CB.
Therefore, the interplay of FP-defects generation, facilitated by electron injection, results in CF formation, which is referred to as the forming step.
Once that this phase took place in the memory cell, the ON-OFF switching procedure can be accomplished for 2 values of polarization: V_{set} and V_{reset}, both of which are lower in magnitude than the forming one[34]. This last statement is elucidated by the fact that, since the CF has already been built in the configuration phase, just a portion of the filament needs to be reconstructed or disrupted to go to LRS or HRS, respectively.

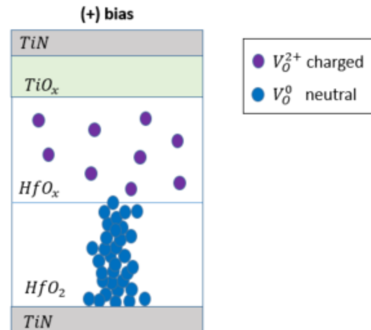


Figure 21: Sketch of CF formation during the SET phase, aided by V_O^{2+} drift from anode to cathode, under the electric field.

2.4.3 Vacancy-assisted CF formation

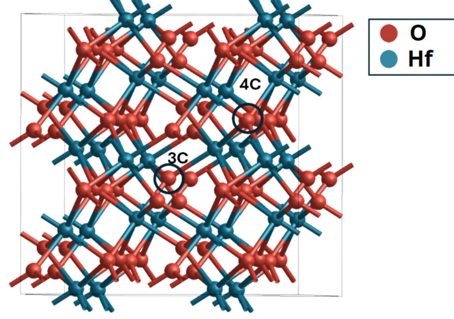


Figure 22: $m - HfO_2$ supercell. Two types of vacancies are found in $m - HfO_2$: threefold (3C) and fourfold-coordinated (4C)

The unit cell of monoclinic HfO_2 is made of 12 atoms: 4 sevenfold-coordinated Hf atoms, and 8 O atoms (4 threefold and 4 fourfold-coordinated).
According to the DFT-studies conducted by *Duncan et al.* in Ref.**[10]**, three possible charge states for the vacancy can coexist within the oxide:

- Neutral V_O^0
- Singly charged V_O^+
- Doubly charged V_O^{2+}

Whether a specific V_O is present in a precise charge state is dictated by the electrode Fermi levels, and this correlation is outlined by the oxygen defects formation energy equation[10]:

$$E_{Form}(q) = [(E(defect, q) - E(bulk, q)] - [E_V(defect, q) - E_V(bulk, q)] \\ + q \cdot (E_F - E_V) + E_{corr} + \sum n_i \mu_i \tag{2.1}$$

where:

- $E(defect, q)$ and $E(bulk, q)$ are the total energies of the defective and pristine crystals having charge state q;
- $E_V(defect, q)$ and $E_V(bulk, q)$ are the valence band energy shifts;
- E_{corr} is a correction energy parameter;
- n_i is the number of oxygen vacancies with charge i, and μ_i is the corresponding chemical potential.

The graphical representation of the formula is depicted in Fig.23, where the formation energy of isolated V_O and clusters is reported as a function of the electrode Fermi level.

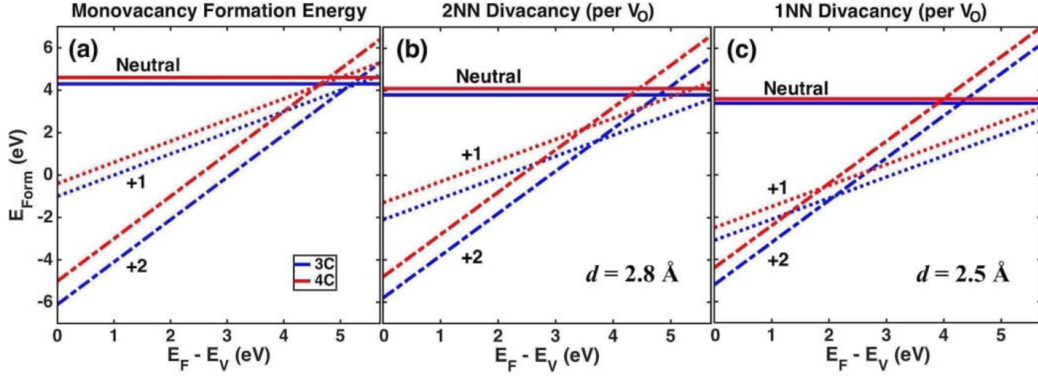


Figure 23: Formation energy vs $E_F - E_V$, for (a) monovacancy, (b) 2NN divacancy (with $V_O - V_O$ distance $d = 2.8\mathring{A}$) and (c) 1NN divacancy (with distance $d = 2.5\mathring{A}$). Adapted from **[10]**

The specific charge state has an influence on the final shape of the filament. Indeed, as illustrated by Fig.23a)., isolated vacancies typically exist in the +2 state, because to them is assigned the lowest E_{form}, for most of the Fermi level range.
However, when a neighbour is introduced, the +1 state is more favorable than the +2 close to the conduction band (Fig.23 b, c.), due to Coulomb repulsion between two V_O^{2+}.
Lastly, when a cluster of oxygen vacant defects is considered, it is said to bring no charge.
At this stage, it is immediately clear that V_O^+ have the tendency to coalesce and form clusterized defective regions, which are the basis of the conductive filament of the memory cell. While V_O^{2+} tends to stay isolated and contribute to the dissolution of the filament.
Both play a crucial role during the set and reset processes, as the V_O^{2+}, being charged, drift from the anode (+V) to the cathode (-V), along the direction defined by the electric field, inside the substoichiometric oxide region. On the contrary, near the cathode, associated to a higher Fermi level, vacancies neutralize and aggregate into filaments (Fig. 21).
This essentially clarifies that despite the presence of closely packaged vacancies, Coulomb repulsion is absent, thereby ensuring the stability of the conductive filament.
Furthermore, the charge state and formation energy of a vacancy are strictly related to the kinetics of V_O's itself. Indeed, the latter is defined by the migration energy barrier of the particle: the lower it is, the more favorable is its diffusion. The *ab initio* energy values for the three charge states are given in figure 24, but in the KMC model (Chap 3), these values will be adjusted to observe their effect.

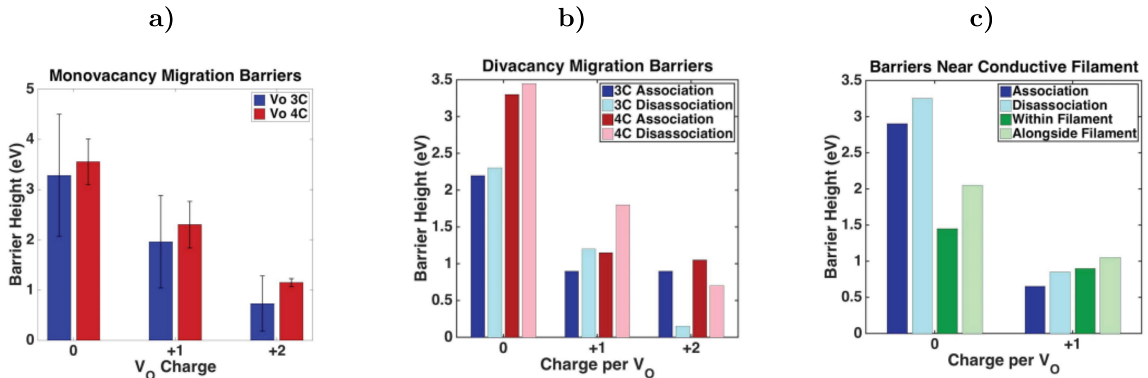


Figure 24: a) Migration energy barrier for 3C, 4C V_O^0, V_O^+ and V_O^{2+} b) Migration energy barrier for association/dissociation of divacancies c) Energy barrier for V_O^0 and V_O^+ migration near and inside the filament. Adapted from **[10]**

Based on the charts presented in Fig.24, three key observations can be made:

- The diffusion of isolated V_O^{2+} is linked to a lower energy barrier, thus making its migration more favorable compared to neutral or +1 vacancies;
- Neutral and +1 V_O are prone to clustering due to a lower energy barrier. Conversely, the +2 charge state tends to dissociate easily due to a more favorable energy landscape for dissociation.
- V_O can also move parallel to the CF axis, within or adjacent to the filament, exhibiting a significant anisotropy of V_O mobility. This depends on the preferential crystallographic direction and the presence or not of grain boundaries (GB) within $m - HfO_2$.

Kinetic Monte Carlo framework

The objective of this Chapter is to initially provide an overview of modeling and the justification behind the use of distinct types of models based on their intended applications.
Specifically, this work concentrates on the KMC method, examining the construction of its framework by investigating subsequent computational modules including the Poisson equation Solver, Charge Transport mechanism, and Heat Flow equation solver. The retention measurements discussed in the next Chapter are conducted using this model as a foundation.

3.1 ReRAM modeling overview

Although significant achievements has been attained in recent years, RRAM technology still has to overcome several challenges in order to meet the requirements summarized in Table 2.

Parameter	Technology requirements
Operating voltage	< 1 V
Power consumption	≈ 10 pJ per transition
Switching time	< 10 ns per transition
Endurance	$> 10^9$ cycles
Data retention	> 10 years
MIM cell size	576 nm^2
I_{ON}/I_{OFF} ratio	10^6

Table 2: Technology requirements for RS-based NVMs. Adapted from [**43**]

Within this context, a well performing RRAM device should exhibit low spatial (cell-to-cell) and temporal (cycle-to-cycle) variability. This implies that each memory cell within the crossbar array should perform in a uniform manner throughout the R/W cycles.
Managing this is complex, as the stochastic nature affects the set/reset transition, leading to CF morphology variations and experimental spread resistance distributions both for LRS and HRS.

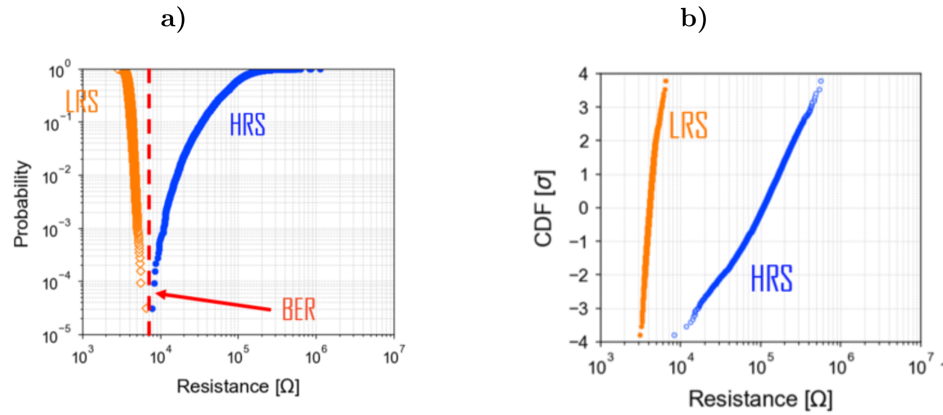


Figure 25: a) Example of LRS-HRS distribution b)Example of QQ plot distribution.

In this uncertain scenario, RS simulations can serve as a valuable tool for handling stochastic effects and interpreting experimental data, as well as for predicting long-term performance by means of the *accelerated life test* (ALT) technique.
Thereafter, once that promising results are obtained for a certain configuration, an optimization of the material stack and array arrangement can be engineered.

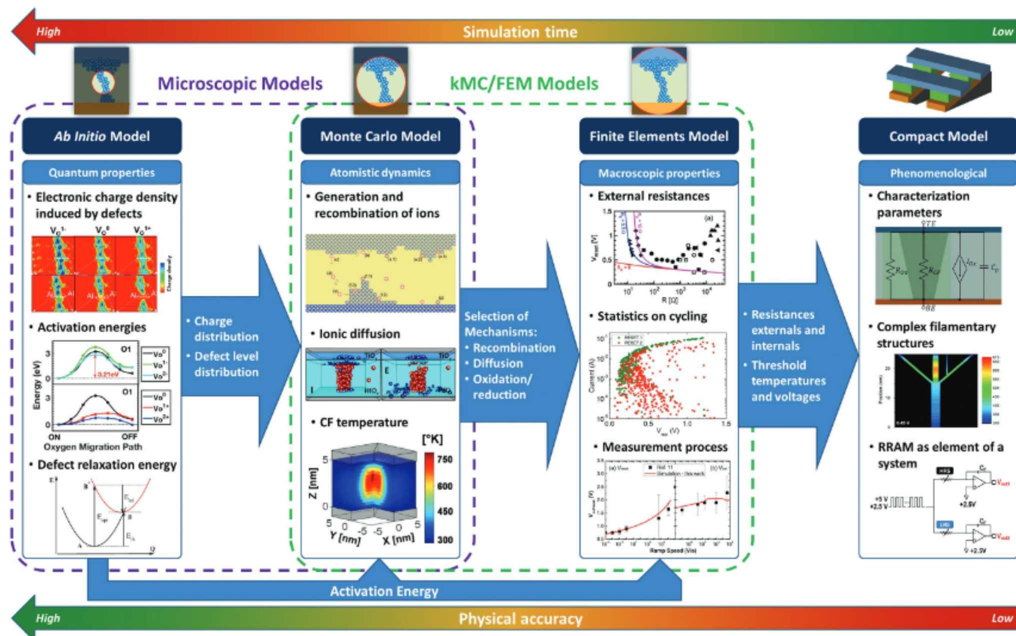


Figure 26: Schematics of the simulation models used for RRAM device characterization. From left to right, the computational time decreases, at the expenses of reduced physical accuracy. Adapted from **[43]**

Such simulation lays its foundation on a model, which could be of various nature, depending on the scope of the calculations to be performed. As a matter of fact, there is not an universal model winning among all, while a hierarchical ecosystem has been developed.
Here, three macro areas can be individuated.

- Material-level atomistic calculations, including *Ab initio* simulations;

- Device models like Kinetic Monte Carlo (KMC) and Finite Element Method (FEM);
- Compact models, running in a SPICE-like environment.

This comprehensive approach attempts to provide a complete understanding of the RS mechanism, whose insights span from fundamental physical properties (such as migration energy barriers for O_i/V_O species), to specific device features and circuit functionalities.
As can be seen in Fig. 26, an increase in physical accuracy will result in higher computational costs. Therefore, it is crucial to select the appropriate method based on the specific requirements.

3.1.1 Microscopic models

Microscopic atomistic models are employed to address the physical process involved during RS within few nm^3 of a filamentary structure such OxRAM[45].
Here, the formation and rupture of CF is facilitated by the generation, recombination, and diffusion of O_i/V_O pairs.
Consequently, electrical, thermal and optical evolution of the material stack is studied, and the charge transport mechanisms are derived.
These heavy calculations are supported by *ab initio* models, including Density Functional Theory (DFT), and molecular dynamics models, together with some experimental insights given as inputs. Examples of the measurements have been already provided in Section2.4.

3.1.2 Device TCAD models

While first principle calculations aim to simulate the physics of switching at the atomistic level, they consider the interplay between just few actors within the cell. In contrast, the dynamics of the device during operation, and its impact on scaling[1] and reliability, is well described by means of technology computer-aided design (TCAD) techniques. Furthermore, in order to address the cell-to-cell variability, stochasticity of events must be considered.
TCAD models can be classified into two main categories[45]: FEM and KMC modeling.

Finite element method (FEM) modeling

In a FEM model, the device volume is discretized into a 2D or 3D mesh of finite elements, where three partial differential equations (PDEs) are solved numerically. These must be solved in a self-consistent way and involve drift-diffusion transport equations for ion migration, carrier continuity equation for electronic conduction, and Fourier equation for Joule heating[45].
Therefore, modeling a RRAM is somewhat more complex compared to, for instance, a *p-n* junction or a CMOS device, because one must consider ionic effects, which are crucial for transport equation. As mentioned earlier, the set/reset process is governed by ionic migration, which can be either:

- Diffusion, occurring in absence of an electric field, driven by concentration and temperature gradients;
- Drift, where an induced electric field F lowers the hopping barrier, causing ions to move as a consequence.

According to this picture, the total ion current density j_D is given by the sum of the two factors[45]:

$$j_D = j_{diff} + j_{drift} = -D\nabla n_D + \mu F n_D \quad (3.1)$$

[1] the geometry to be studied is of the order of magnitude of $100nm^3$ [45]

where D $[cm^2/(V \cdot s)]$ is the ionic diffusion coefficient, n_D $[cm^{-3}]$ is the ion concentration, and μ $[cm^2/(V \cdot s)]$ is the ionic mobility. Note that diffusivity and mobility depends on temperature according to the Arrhenius and Einstein laws, respectively[45]:

$$D = D_0 e^{-\frac{E_A}{k_B T}} \qquad \mu = \frac{qD}{k_B T} \tag{3.2}$$

with E_A the activation energy barrier of the potential well (Fig.39)
Due to O^- and V_O migration the local resistivity of a restricted area of the active material, that is to say the CF, changes. Hence, the filament can be thought of as a highly doped portion of the active material, where oxygen vacancies and hafnium ions act as doping centers.

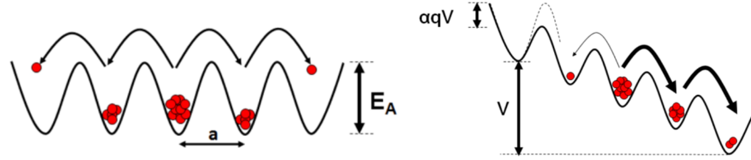


Figure 27: Hopping mechanism described by (left) ion diffusion due to temperature and concentration gradients (right) ion drift due to applied E-field. Adapted from **[45]**

Kinetic Monte Carlo (KMC) modeling

An alternative TCAD approach for simulating the device physical operation and its reliability with time is KMC. Unlike FEM, which handles continuous quantities, KMC considers the individual contributions of ions and vacancies by using discrete variables[45].
Alongside to defect migration, two other mechanisms are included in RS for KMC simulation: generation and recombination of O_i/V_O pairs. Particularly, generation is the main cause behind the forming process, while recombination accounts for the reset operation. These mechanisms, and their relevant formulas, will be elucidated in more detail in Par. 3.2.2.
The variation in resistance is attributed to *trap assisted tunneling* (TAT) as charge transport mechanism, whereby electrons move through percolating paths between traps, aligned in a sequence according to the filament shape, from the cathode to the anode. The disposition and concentration of V_O determine whether the cell will be in LRS (Fig.28b) or HRS (Fig.28a,c). A detailed explanation of TAT can be found later in Par 3.2.5.
The KMC method aligns well with experimental results; in particular, a higher V_{reset} typically leads to an increased gap and final resistance, whereas a larger compliance current results in higher vacancy concentration and, consequently, lower resistance.

a) **b)** **c)**

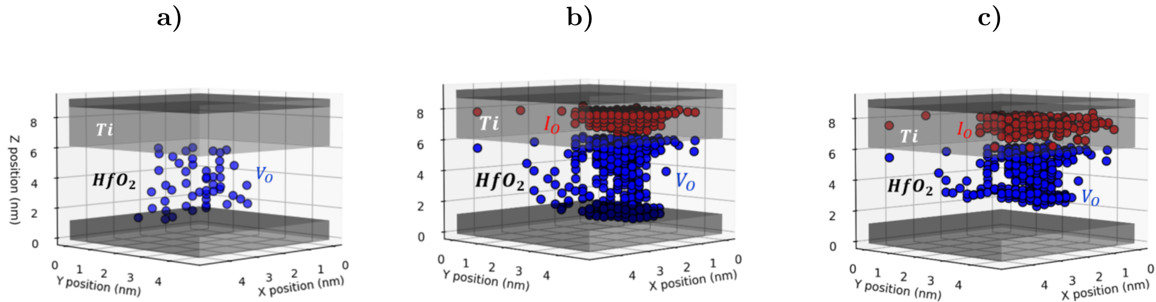


Figure 28: 3D representation of a single OxRAM cell with KMC method. In blue are reported the oxygen vacancies, in red are outlined the oxygen ions. a) Pristine state. b) Low resistance state, after forming. c) High resistance state, after reset programming.

3.1.3 Semi-empirical models

Compact modeling comes into play when large-scale simulations must be conducted, to study analytically device-circuit-system behavior. For instance, at *Weebit Nano*[37], 1T1R memory arrays having storage density from 16kb up to 1Mb are modeled.
Usually, these simulations run in a SPICE-like environment, properly adjusted through FEM/KMC inputs, to assess intrinsic cell variability, and some empirical formulas, derived from experimental measurements.
Compact models are essential to fill the gap between single device dynamics and circuit performances, so as to give hint for its design and optimization.
By taking into account also the parasitic effects, like electrode capacitance C_P and contact resistance R_H, and the leakage paths R_P in the oxide layer, the RRAM cell can be modeled with the following circuit.

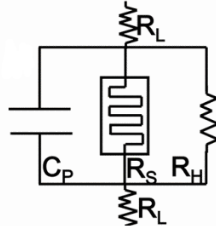


Figure 29: Circuit model of the RRAM cell and its parasitic elements. Adapted from [**47**]

3.2 KMC framework

The objective of this work is to assess the retention and reliability of an OxRAM device by means of a Kinetic Monte Carlo model. The ensuing sections will provide a detailed explanation of the employed multi-physics KMC framework, comprised of several computational modules, as reported in Fig30.

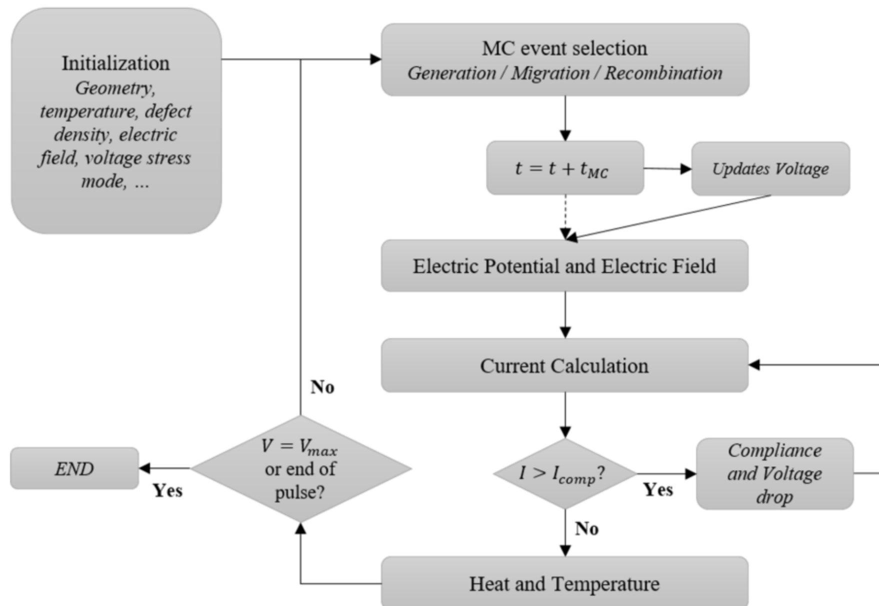


Figure 30: KMC workflow. Adapted from [**44**]

3.2.1 Parameters Initialization

The KMC model must be initially fed, directly at the user interface, by some "realistic" parameters, extrapolated from *ab initio* calculations. These inputs are said to be:

- Geometry, in terms of width and size of the memory cell and number of mesh units;
- V_O defect density;
- Temperature;
- Electric field, so forming, set and reset voltage.

Geometry

The OxRAM cell is realized though a 3D *numpy array* simulating the TiN/Ti/HfO$_2$/TiN MIM stack. Therefore, a simplified version of a real memory cell was utilized as a reference since, in practice, many more layers and variations of the MIM stack exist, along with the doping of Hafnia.
Furthermore, the dimensions itself were shrink to 5nm×5nm×10nm cell, which is larger than the $100nm^3$ scale of Ref[**45**], but different from the 10nm×10nm×30nm experimental cell.
The simulated domain was discretized with cubic mesh unit of size $dx = 2.6\mathring{A}$, similar to the interstitial $O^{2-} - O^{2-}$ and $V_O^{2+} - V_O^{2+}$ distance in $m - HfO_2$ according to first-principle calculations of Ref. [**48**].
The final mesh is structured and uniform, with well-defined number of nodes N_x, N_y, N_z that physically represent the thickness of each layer within the device.

```
Nz = N_elec_top + N_Tiox + N_ox + N_elec_bot
shape = (Nz, Ny, Nx)
```

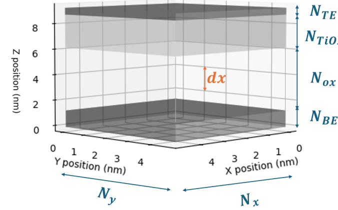


Figure 31: Mesh of the OxRAM cell, with step size $dx = 2.6\mathring{A}$. The top and bottom electrodes are represented in dark grey; the scavenging TiO_x in grey; and the active HfO_2 layer in white.

Defect density

Within the 3D grid, vacancies are restricted to move within the active medium, while ions can cross the Ti layer, which, due to its relatively low electronegativity, can be oxidized and form a non-stoichiometric layer of TiO_x [10][33][34], as previously mentioned in Par.2.4.1.
During the forming phase, an initial V_O^{2+} defect density was randomly assigned to the oxide layer according to a V_O fraction probability, set by the user.

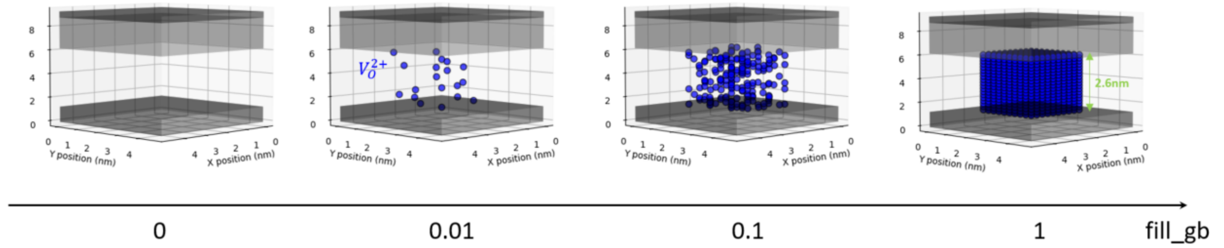


Figure 32: Effect of the *fill_gb* parameter on the Hafnia defect density. Being a probability, *fill_gb* varies from 0 (left, monocrystal oxide) to 1 (right, anisotropic oxide). The defective portion can be properly adjusted. In this case, an inner cubic region of size 2.6nm can be filled with vacancies.

A remarkable point to be underlined in this area is that, when launching a new Forming simulation, the resulting V_O density and positions vary even if the user input parameters (e.g., fill_gb=0.01) remain constant. This aligns with the stochastic nature of KMC simulations, which depend on event probabilities. Hence, this inherent randomness is exploited to perform multiple experiments, providing a distribution of resistance values.
To accurately assess the behavior of a specific programming condition, it is necessary to run hundreds of simulations and derive statistical measures, as a single KMC simulation is meaningless on its own.

Temperature and applied voltage

The defined geometry is also fundamental for the calculation of heat and electric field maps, accomplished through Finite Difference Method (FDM). For this reason, heat P, temperature T, permittivities ε_r, voltage V and electric field (E_x, E_y, E_z) were initialized with the same *shape* dimension of the grid.
In the following is reported the 2D $(z - x)$ contour plot of the electrostatic potential, when a bias of +1V is applied on the top electrode ($z = 9.36nm$) of an OxRAM cell in pristine state.

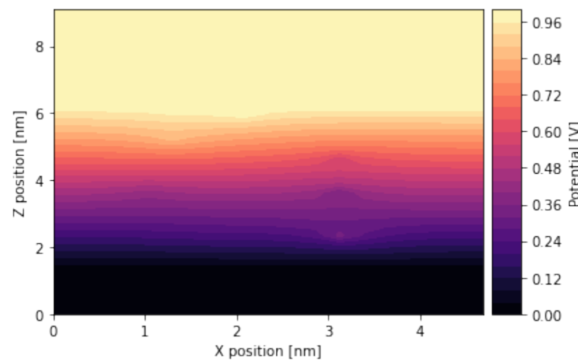


Figure 33: 2D Potential map in the $z-x$ plane. A bias voltage of +1V is applied at the top electrode. The electrostatic potential lines distort in correspondence of the vacancy sites.

During the initialization step of the KMC module, depending on the type of operation to be performed on the memory, either the forming, set or reset voltage, V_{DD}[2] is properly adjusted in terms of magnitude and pulse width.

[2]voltage applied on the 1T1R stack

According to Ref.[**9**], the larger is the pulse magnitude and/or width of V_{reset}, the larger will be the gap and so the higher will be the resistance.
On the contrary, for higher compliance current in the set programming, the number of percolation paths within the filament is increased, making the overall resistance lower and, hence, the LRS more stable.

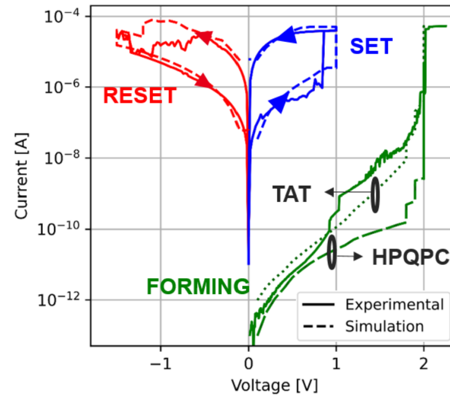


Figure 34: Typical bipolar I-V characteristics including Set, Reset and Forming operations, derived both from experiments (solid line) and simulations (dashed line). Non-linearity during Forming is ensured with MTAT Solver. Adapted from [**46**]

Parameter	Description	Value
d	$V_O^{2+} - V_O^{2+}$ distance	$2.6\AA$
N_{TE}	Top electrode z-thickness	2
N_{BE}	Bottom electrode z-thickness	5
N_{FE}	Floating electrode z-thickness	2
N_{OTS}	OTS z-thickness	38
N_{TiO_x}	TiO_x z-thickness	10
N_{HfO_2}	HfO_2 z-thickness	19
N_x	Thickness along x	19
N_y	Thickness along y	19
u_T	Top electrode voltage	
u_B	Bottom electrode voltage	
u_F	Floating electrode voltage	
GB_d	Grain boundary diameter	3nm
N_{grain}	Grain initial defect concentration	$1 \times 10^{20}/m^3$
N_{GB}	GB initial defect concentration	$2 \times 10^{20}/m^3$
$fill_{GB}$	GB initial defect probability	0.02
T_{amb}	Ambient temperature	300K
h_x	Heat exchange coefficient along x (HfO_2 walls)	4×10^9 m/s
h_y	Heat exchange coefficient along y (HfO_2 walls)	4×10^9 m/s
h_z	Heat exchange coefficient along z (TiN walls)	80×10^9 m/s
k_{th,HfO_2}	HfO_2 thermal diffusivity	$9.5 \times 10^{-7} m^2/s$
$k_{th,Ti}$	Ti thermal diffusivity	$2.6 \times 10^{-6} m^2/s$
$k_{th,TiN}$	TiN thermal diffusivity	$9.3 \times 10^{-6} m^2/s$
$E_{A,G}$	Hf—O bond breakage zero-field activation energy	2.8eV
$E_{A,R}$	O_i/V_O pair recombination migration energy	0.2eV
$E_{A,D}$	Migration activation energy	V_O^{2+}: 1eV V_O^{1+}: 1.1eV V_O^0: 2eV O^{2-}: 0.6eV
b	Hf—O bond polarization factor	0.4
ν_0	Effective Hf—O bond vibration frequency	7×10^{13} Hz
L	Lorentz factor	1/3
χ	Amorphous HfO_2 electric susceptibility	15
ε_{r,HfO_2}	HfO_2 relative permittivity	16
ε_{r,TiO_x}	TiO_x relative permittivity	80
$\varepsilon_{r,metal}$	TiN and CF relative permittivity	1024
m^*	Electron effective mass	$0.18 \times m_e$
I_{CC}	Pulse generator compliance current	
nmos type	Series transistor type	
V_G	Gate voltage of the series transistor	
V_{read}	Read voltage	
V_{max}	Programming pulse voltage	
V_{step}	Voltage step of voltage pulse ramp	0.1V
$t_{ramp,up}$	Programming pulse ramp up time	
$t_{ramp,dw}$	Programming pulse ramp down time	
$t_{plateau}$	Programming pulse plateau time	
t_{WT}	Write termination trigger time	
$n_{actions}$	Number of KMC actions per cycle	5

Table 3: Simulation initialization main parameters to be tailored by the user. In this work, floating electrode and OTS selector are not considered.

3.2.2 Generation, Recombination and Migration

After initializing all the variables, the simulation enters in the main loop of the process, the one regarding event selection according to the Monte Carlo method.
Before delving into the description of this mechanism, first the physical meaning of events and their corresponding transition rates are discussed.
An analysis is conducted on 14 potential events per each mesh point, which encompass O_i/V_O generation, recombination, and their diffusion along x, y, z, in line with the Frenkel-pair theory elaborated in Par. 2.4.2. This section will outline the formulae based on Transition State Theory, implying that transition rates can be viewed as the frequency (measured in s^{-1}) at which each specific event occurs. In turn, the rates follow an Arrhenius-like formulation, where the higher is the activation energy E_A, the more the reaction is sensitive to a temperature change:

$$\Gamma = A \cdot \exp\left(-\frac{E_A}{k_B T}\right) \quad \longrightarrow \quad \ln\Gamma = \ln A - \frac{E_A}{k_B T} \tag{3.3}$$

- Concerning doubly charged oxygen ion and vacancy, the generation rate of a V_O^{2+}/O^{2-} FP as a result of the Hf-O bond breakage can be modeled as follows[46][44].

$$\Gamma_G(x,y,z) = \nu_0 \ \exp\left[-\frac{E_{A,G} - bE(x,y,z)}{k_B T(x,y,z)}\right] \tag{3.4}$$

  where E and T are the local electric field and temperature derived from the solution of Poisson and Fourier heat's equation (Par.3.2.7), while b is the $Hf - O$ bond polarization factor and k_B is the Boltzmann constant.
  This formula underlines that as the electric field increases, Γ_G also increases, thereby making generation more favorable to occur, since the barrier of $E_{A,G} = 2.8eV$ is reduced.

- Much lower barrier energy (0.2eV) characterizes, on the contrary, the V_O^{2+}/O^{2-} FP recombination process, which depends on the local temperature within the active medium. The corresponding transition rate reads[46][44]:

$$\Gamma_R(x,y,z) = \nu_0 \ \exp\left[-\frac{E_{A,R}}{k_B T(x,y,z)}\right] \tag{3.5}$$

- Finally, it is fundamental to characterize the atomic diffusion of V_O^{2+} and O^{2-} in order to address RS phenomenon. Here, each atomic species can migrate in six possible directions: left, right, up, down, forward, and backward.
  The diffusion rate equation implemented in the model is[46][44]:

$$\Gamma_D(x,y,z) = \nu_0 \ \exp\left[-\frac{E_{A,D} - \alpha Q d(1 + L\chi)E}{k_B T(x,y,z)}\right] \tag{3.6}$$

  where L and χ are defined in Table 3, and $Q = \pm 2$ depending if it deals with oxygen vacancies or ions. The electric field, instead, could be E_x, E_y or E_z, for x, y or z-oriented migration.
  Finally, α represents the symmetry factor of the state transition which, in this work, was set to 0.5. That is indicative of an energy barrier placed symmetrically in between initial and final state.

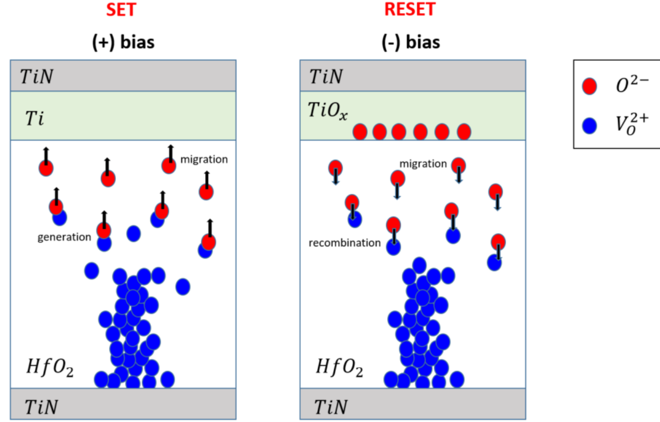


Figure 35: Generation and recombination of O^{2-}/V_O^{2+} pairs. The first is typical of Set operation (positive bias at the TE), while the second is more likely to occur during Reset programming (negative bias at the TE). Ions move accordingly to the direction of the electric field.

3.2.3 Monte Carlo event selection

Once having presented the three physical process ruling RS in an OxRAM, the system is ready to implement the Monte Carlo method. To this aim, the KMC workflow proceeds with the estimation of the occurrence probability per each event, the calculation of the next iteration time and, ultimately, the selection of the winning outcome.
In this particular version, the system has no memory of what happens the iteration before, meaning that each O_i/V_O state transition is independent from the previous one. The reason for that has to be found in the fact that the time for a transition to take place is shorter than the actual waiting time for the action.
In formulaic terms, this is expressed with an exponential decay, where the probability P of a specific event to take place after a time $t_{\text{iteration}}$ depends on the particular event rate according to[46][50]:

$$P = 1 - \exp(-\Gamma \cdot t_{\text{iteration}}) \tag{3.7}$$

Thus, the longer the time, the more probable is the transition to occur.
Consequently, the probability of nothing to happen can be derived easily[46][50]:

$$P_{\text{none}} = 1 - \prod_{\text{atoms \& ions}} P = \exp\left(\sum_{\text{atoms \& ions}} \Gamma \cdot t_{\text{iteration}}\right) \tag{3.8}$$

In this study P_{none} was set to 50% as a trade-off between computational expense and precision[46][50].
From eq.3.8 the next KMC iteration time is inversely proportional to the cumulative sum of all the occurring event rates as:

$$t_{\text{iteration}} = -\ln(P_{\text{none}}) \cdot \sum_{\text{atoms \& ions}} \Gamma \tag{3.9}$$

The latter is incremented with the running physical time t, and after selecting the winning event, the global time is updated as:

$$t = t + t_{MC} \quad t_{MC} \equiv t_{\text{iteration}} \tag{3.10}$$

To prevent multiple transitions from happening in the same iteration for any single particle, the probability of each event has been normalized between 0 and 1.
Then, in line with the KMC procedure, a random number is picked within the same range per each

particle, and the transition probability corresponding to this random number is selected as winning event, discarding all the others.

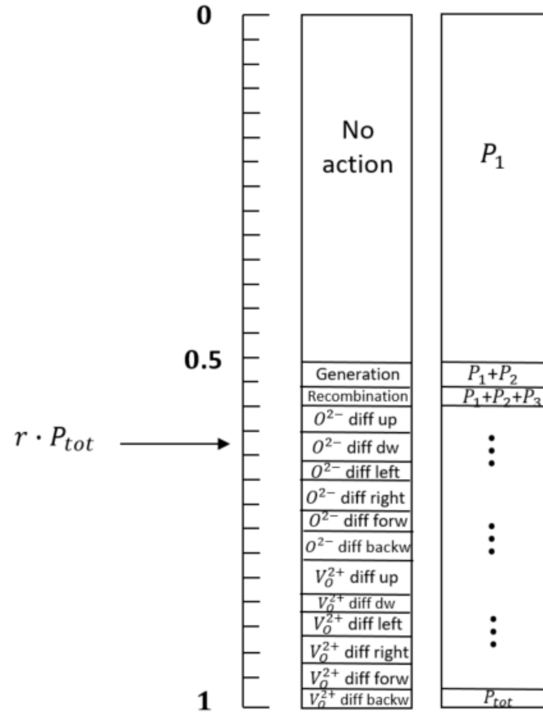


Figure 36: Diagram illustrating the Monte Carlo event selection process. Here, the random number lies between 0.35 and 0.4, which corresponds to the probability of the downward ion migration. As can be observed by the entity of the boxes, each event has a different occurrence probability, because each transition rate is calculated.

3.2.4 Electric Field and Electric Potential Poisson Solver

The Electric Field is fundamental for the formation or dissolution of the CF inside the memresistive layer, since it acts as an electromotive force for the migration of the charge species.
For each iteration of the KMC loop, the electric field distribution is calculated through a Poisson Solver, aimed to find a numerical solution of the *3D-generalized Poisson equation*, leveraging on FDM techniques.[44]

$$\nabla \cdot [\varepsilon_r(\mathbf{r})\nabla V(\mathbf{r})] = -\frac{\rho(\mathbf{r})}{\varepsilon_0} \quad \text{or} \quad \nabla^2 V(\mathbf{r}) = -\frac{\rho(\mathbf{r})}{\varepsilon_r \varepsilon_0} \tag{3.11}$$

In terms of electric field it can be rewritten as:

$$\nabla \cdot [\varepsilon_r(\mathbf{r})E(\mathbf{r})] = \frac{\rho(\mathbf{r})}{\varepsilon_0} \tag{3.12}$$

with ρ the free charge density, ε_0 the vacuum dielectric constant, and $\varepsilon_r(\mathbf{r})$ the relative permittivity map of the specific layer[3] to be updated each iteration, according to the oxygen ion and vacancy coordination number.
This last passage is crucial, since, just to mention, in presence of an oxygen vacancy, the conductivity of Hafnia is locally increased. Therefore, the direct consequence is that, when the vacancy concentration

[3]either TiO_x or HfO_2

gets higher, a localized metallic nature can be observed within the active medium, allowing the electric field to concentrate in the proximity of vacancy clusters.
Returning to the Poisson equation, obtaining a closed-form analytical solution for 3.11 is unfeasible, necessitating the use of a numerical solver. The elective technique was the Finite-Difference Method (FDM), where $\{N_x\}_{i=1}^{N}$, $\{N_y\}_{i=1}^{N}$, $\{N_z\}_{i=1}^{N}$ are, respectively, the x, y, z nodes of the 3D regular mesh which constitute the memory cell.
Referring, for simplicity, to the vacuum case, the extended version of eq.3.11 is[46]:

$$\frac{\partial^2 V(n,m,k)}{\partial x^2} + \frac{\partial^2 V(n,m,k)}{\partial y^2} + \frac{\partial^2 V(n,m,k)}{\partial z^2} = -\frac{\rho(n,m,k)}{\varepsilon_0} \tag{3.13}$$

Then, the central difference approximation is applied to each partial derivative[46].

$$\frac{\partial^2 V(n,m,k)}{\partial x^2} \approx -\frac{V(n-1,m,k) - 2V(n,m,k) + V(n+1,m,k)}{h^2} \tag{3.14}$$

with h the mesh grid spacing.
By substituting 3.14 into 3.13, Poisson equation reads[46]:

$$\begin{aligned} V(n-1,m,k) + V(n+1,m,k) + V(n,m-1,k) + V(n,m+1,k) + \\ V(n,m,k-1) + V(n,m,k+1) + 6V(n,m,k) = \frac{h^2}{\varepsilon_0}\rho(n,m,k) \end{aligned} \tag{3.15}$$

Along with this, boundary conditions have to be defined at the edges of the 3D volume.
In particular, at the top and bottom electrodes, along z, the choice fell into *Dirichlet Boundary Conditions* (DBC), while for the lateral x, y walls *Neumann Boundary Conditions* (NBC) were preferred. Regarding the latter, the gradient of the electric potential at the lateral domain edges was set to zero, to ensure continuity of the electric field at the boundaries.
At the end, when considering a non-uniform dielectric medium, the Poisson equation can be synthesized into a system of linear equations having the form:

$$Ax = b \tag{3.16}$$

where A is a 3D matrix containing the $a_{n,m,k}$ constants which depends directly on the local permittivities. Then, the second member is equal to $b = -\frac{h^2}{\varepsilon_0}\rho(n,m,k)$. Finally, x represents the 3D potential, to be calculated by means of the inversion technique.
From this, the electric field map is easily obtained as its derivative.

3.2.5 Current calculation and conduction mechanism

At this stage of the KMC workflow, the charge transport model plays a critical role in determining how the CF configuration switches between HRS and LRS.
Within the KMC framework two current solvers were implemented: one based on *Trap Assisted Tunneling* (TAT) and the other relying on the creation of *Quantum Point Contact* (QPC) resistances. The first is used to describe the system behavior for comparably low V_O concentrations (said to be during Forming), while QPC is exploited for higher V_O concentrations (Set, Reset and Relax).

- The starting point for the definition of the multiphonon TAT process is to define what is meant by "traps". They are localized defective states within the bandgap of the dielectric medium, specifically known as Vacancies in this scenario.
  Traps are able to capture electrons from the conduction band of the cathode. Electrons, in turn, tunnel through the dielectric potential barrier and reach the anode. The emission of an electron from a trap, makes it ionized, and broadens the energy level E_T. However, to reach the

anode conduction band, some additional energy is required. The latter is ensured by the electron coupling with multiple oxide phonons, which can be thought as quantized modes of vibration occurring in the lattice. The energy carried by a single phonon is equal to $\hbar\omega_0 = 0.07eV$, thus the final amount of provided energy is $E_{\text{capt}} + m\hbar\omega_0$.
MTAT via positively charged oxygen vacancies is essential for the conduction during the Reset phase[51]. In fact, when in HRS, the CF is partially ruptured, but since clusters of traps are still present, these can help to transfer charge through the, even if incomplete, *percolation paths.* The creation of a larger number of percolation paths, then, allows for more charge to be transferred and, therefore, for the transition to LRS.

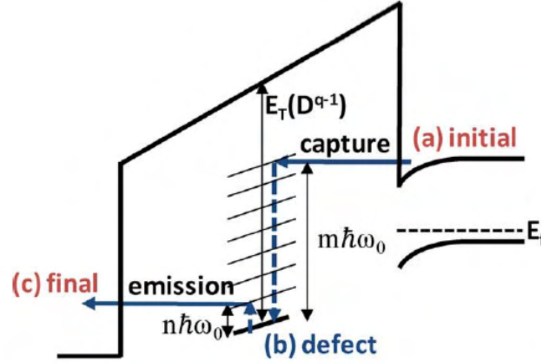


Figure 37: Schematics of multiphonon TAT for a single-trap percolation path. The electron is initially captured (a) from TiN CB at the Trap level E_T. During this phase, the electron crosses through virtual states (b), aided by multi-phonon energy $m\hbar\omega_0$. Finally, the electron is emitted at the Ti CB (c). Adapted from **[51]**

A percolation path is made of n-traps. In order to determine the current flowing, the rate R of electrons passing through the n-traps was calculated by the TAT Solver.
In particular, being $R_{c,j}$ and $R_{e,j}$ the rate at which an electron is captured and emitted by and from trap j[46]:

$$R_{c,j} = \tau_{c,j}^{-1} \cdot (1 - f_{t,j}) \tag{3.17}$$

$$R_{e,j} = \tau_{e,j}^{-1} \cdot f_{t,j} \tag{3.18}$$

these can be assumed to be equal under steady-state condition: $R_{c,j} = R_{e,j}$. As a consequence, the occupation probability of the trap is $f_{t,j} = 1/2$. Whereas, the rate at which charge passes through trap j depends only on the time constants $\tau_{c,j}$ and $\tau_{e,j}$[46].

$$R_j = R_{c,j} = R_{e,j} = (\tau_{c,j}^{-1} + \tau_{e,j}^{-1})^{-1} \tag{3.19}$$

In addition to this, since the emission rate at trap j is equal to the capture rate at trap $j+1$:

$$R_{c,j+1} = R_{e,j} \quad \longrightarrow \quad R_{j+1} = R_j \tag{3.20}$$

the conclusion is that the rates at which electrons passes through all the n traps of the percolation path are equal.
In other terms, in a percolation path made of multiple similar traps, the global rate R at which charge flows is limited by the slowest trap[46].

$$R = \frac{1}{\max_i(\tau_{c,j}^{-1} + \tau_{e,j}^{-1})} \tag{3.21}$$

The higher are $\tau_{c,j}$ and $\tau_{e,j}$, the slowest is the rate, affecting the final current.
The objective of the Solver is to individuate the slowest percolation path among all the possible ones, and this was accomplished through a modified Dijkstra Steady State Percolation Adapted Algorithm (DSSPAA)[46].

- As previously mentioned, there is, actually, a second Solver which was implemented in the model. The explanation behind this choice is that the MTAT percolation algorithm was computationally heavy, and a higher-throughput option was preferred.
Nevertheless, TAT was still used for modeling charge transport in presence of low defect density, such as during Forming operation. On the other hand, when V_O concentration is high, the *Hybrid Poisson QPC Solver* (HPQPC) is enforced. In this scenario, each trap-to-trap link is considered as a quantum point contact and, ultimately, the CF itself can be assessed as a linear network of conductances $G_{V_O-V_O}$ (so resistances). The latters are determined according to Landauer Theory[53]:
$$G_{V_O-V_O} = G_0 \exp\left[2\frac{\sqrt{2m_e^* E_C}}{\hbar}(r_t - a)\right] \tag{3.22}$$
Here, several actors are involved:
  - G_0 is the quantum conductance equal to $7.7510 \times 10^{-5} S$;
  - E_C is the source-trap barrier height needed to be overcame to make electron tunneling possible;
  - m_e^* is the electron effective mass;
  - a is the distance between two V_O;
  - r_t is the trap radius, here assumed to be equal to the mesh unit size $d = 0.26nm$.

  In the present work, only two parameters are variable: E_C and a. To be specific, the activation energy is initially 3.2eV at the electrode interfaces and varies randomly between 1.7eV and 2.7eV within the oxide. Subsequently, it is set to the same value at the electrodes and 1.7eV throughout the oxide.

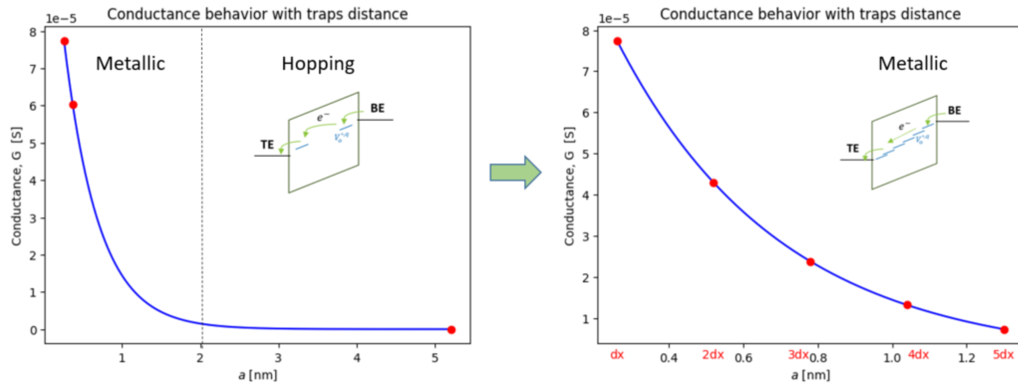


Figure 38: Conductance exponential behavior with trap distance a, according to eq.3.22. For distances higher than 2nm a transition from metallic to hopping conduction is registered. The plot was extracted assuming $E_C = 1.7eV$ in the oxide.

Fig.38 shows that the maximum conductance value, G_0, occurs at the minimum distance $a \equiv dx = 0.26nm$, which corresponds to the mesh unit size. Calculations cannot be performed for

distances smaller than this.
For increasing trap distance, then, the conductivity decreases exponentially and that explain the high resistance in ruptured or dissolved CFs. From a physical viewpoint, what is happening is a transition between a metallic conduction, due to overlapping defect levels, to a hopping conduction, where, as the name suggests, the electron has to "hop" from one trap level to another one.
In the HPQPC solver, the current flowing through the device and the electrical potential map were computed contemporary, to save computational time. With this purpose, a linear system of equation was built, starting from the Kirchooff's law for each trap and electrode[46].

$$\begin{cases} \sum\limits_{i\neq 0} I_i = \sum\limits_{i\neq 0} (V_i - V_0)G_{i,0} = I_{out} = -I_{in} & \text{ANODE} \\ \sum\limits_{i\neq 1} I_i = \sum\limits_{i\neq 1} (V_i - V_1)G_{i,1} = 0 \\ \sum\limits_{i\neq 2} I_i = \sum\limits_{i\neq 2} (V_i - V_2)G_{i,2} = 0 \\ \vdots \\ \sum\limits_{i\neq n} I_i = \sum\limits_{i\neq n} (V_i - V_n)G_{i,n} = 0 \\ \sum\limits_{i\neq n+1} I_i = \sum\limits_{i\neq n+1} (V_i - V_{n+1})G_{i,n+1} = I_{in} & \text{CATHODE} \end{cases} \tag{3.23}$$

Where $i = 0$ and $i = n + 1$ represent anode and cathode, respectively, I_i and V_i are the current and electrical potential at the ith trap, and $G_{i,j}$ is the trap-to-trap conductance of eq.3.22.
The G matrix has dimension equal to the oxide volume and it was built starting from this system, neglecting first and last row[46].

$$\begin{cases} V_1 \sum\limits_{i\neq 1} G_{1,i} - \sum\limits_{i\neq 1} G_{1,i} V_i = 0 \\ V_2 \sum\limits_{i\neq 2} G_{2,i} - \sum\limits_{i\neq 2} G_{2,i} V_i = 0 \\ \vdots \\ V_n \sum\limits_{i\neq n} G_{n,i} - \sum\limits_{i\neq n} G_{n,i} V_i = 0 \end{cases} \tag{3.24}$$

From this, the electrical potential was obtained from any trap[46]:

$$V_i - \frac{\sum\limits_{i\neq j} G_{i,j} V_j}{\sum\limits_{i\neq j} G_{i,j}} = 0 \quad i, j = 0, 1, 2, .., n + 1 \tag{3.25}$$

Lastly, it is worth to mention that conduction is ensured only by trap-to-trap point contacts, implying that, ionic effects are not considered. This is a strong assumption, but it does not have a large impact on the final results.

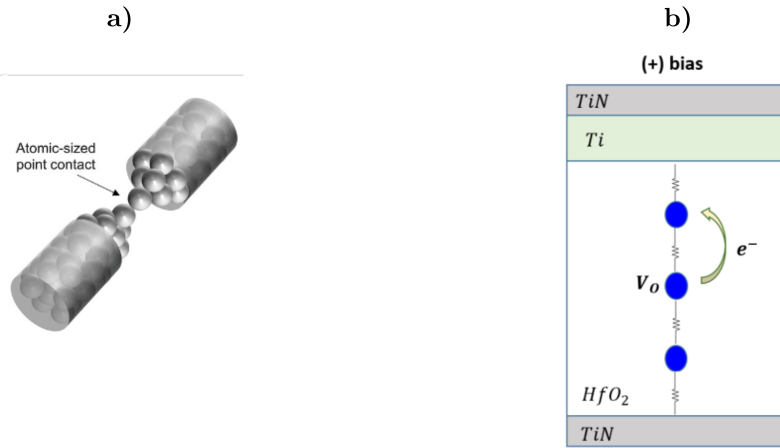


Figure 39: a) Conductance quantization in a metallic inclusion. Adapted from **[52]** b) Schematics of quantum conductance points in the OxRAM cell.

3.2.6 Compliance and Voltage drop

Within the KMC framework, the user can choose whether to apply a compliant transistor to the OxRAM. If applied, the ultimate current compliance will be governed by the series with the n-type MOS.

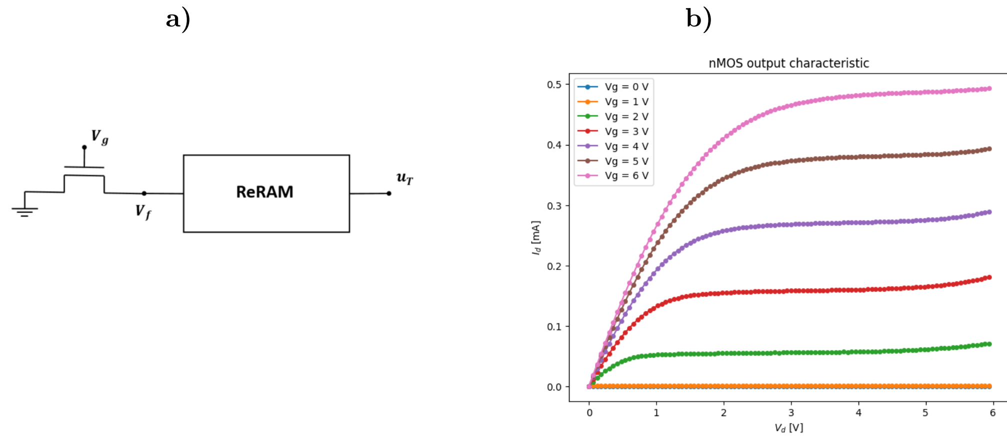


Figure 40: a) Circuital scheme of the 1T1R configuration implemented in the KMC model. u_T is the top electrode applied bias, V_g is the 1T gate voltage and V_f is the floating voltage. b) nMOS $I_d - V_d$ output characteristic for varying V_g.

When dealing with an experimental setup, current compliance was a crucial aspect to be determined, because the series with the MOS prevents a too strong dielectric breakdown. Moreover, in the context of a crossbar architecture, the selector is fundamental to limit the sneak path currents and ameliorate the current ratio between the selected and non-selected memory cells, and, ultimately, ensure a good device reliability.
For this purpose, several 1T, experimentally extracted, characteristics were already present in the model (Fig.40b). Their gate voltage $= V_G$ is fundamental to set the final compliance current.
Referring to the scheme of Fig.40a), to exactly derive how the voltage distributes between the MOS and the load, the current continuity equation was computed.

The scheme is the following[4].

- If $I > I_{CC}$, the voltage drop across the OxRAM is calculated:

$$u_T = V_{dd} \cdot \frac{I_{CC}}{I} \tag{3.26}$$

- If $I < I_{CC}$, the drop of voltage across the MOS is negligible:

$$u_T = V_{dd} \tag{3.27}$$

3.2.7 Temperature update through Fourier Heat equation

Along with Electric Field, in each iteration step, the Temperature map is updated too. Indeed, both the electrical permittivity fluctuations and temperature increase can be addressed as the cause of the CF morphology[44], especially during the Forming dielectric-breakdown phase [54][55].
In this picture, heat generation depends directly from the MTAT process described in Par.3.2.5. As depicted in Fig.37, when an electron is first captured by a trap and then released into the *Ti* CB, there is a net release of energy given by[46][56]:

$$\Delta E_{MTAT} = (m - n)\hbar\omega_0 \quad \text{with } m > n \tag{3.28}$$

The global power dissipation associated to TAT was, then, multiplied by the rate of charge flow through traps (eq. 3.21)[46][56]:

$$P_{MTAT} = R \cdot \Delta E_{MTAT} = R(m - n)\hbar\omega_0 \tag{3.29}$$

Conversely, during the Set and Reset phases, the conduction model used was QPC, with Joule heating being responsible for power dissipation. The latter was determined by adding up all the Joule heating contributions from each of the current inputs, as shown in[46]:

$$P_{QPC_i} = RI^2 = \frac{1}{\sum\limits_{j \neq i} \frac{1}{R_{j,i}}} \left(\sum_{j \neq i} I_{j,i} \right)^2 \tag{3.30}$$

with $R_{j,i}$ and $I_{j,i}$ the resistance and the current between traps j and i, respectively.
Once defining the dissipation map, it is harnessed to determine how temperature evolves within the device. To achieve this, the time-dependent 3D *Fourier's heat flow equation* was computed using the same numerical solver adopted for the Poisson equation[46][44].

$$\nabla(k_{TH} \cdot \nabla T) = P + \rho c_p \frac{\partial T}{\partial t} \tag{3.31}$$

Where thermal conductivity k_{TH}, material density ρ, and heat capacity c_p vary throughout the MIM stack (see Table 3). Whereas the temperature is a local function of space (x, y, x) and time t.
Nevertheless, in nanoscale systems, accurately resolving the temperature is challenging due to the rapid and stochastic nature of the internal processes occurring within the structure. As a result, the switching operation is considered to be an adiabatic process[57], where the obtained temperature values are often indicative and might be overestimated.
In the KMC model, the 3D FDM-based Poisson Solver is adapted for the resolution of equation 3.31. Here, the boundary conditions play a crucial role in the simulation of thermal distribution. In particular, NBC were employed, stating that the heat flux registered at the six domain walls is given by a specified value:

$$\frac{\partial \Phi}{\partial S} = h(T_1^{n+1} - T_{amb}) \tag{3.32}$$

[4]positive bias $V_{dd} > 0$ has been assumed

$$\frac{\partial \Phi}{\partial S} = k_{TH} \frac{\partial T}{\partial x}\bigg|_{1}^{n+1} \tag{3.33}$$

Here, T_{amb} is the ambient temperature, T_1^{n+1} is the temperature at time $n+1$ in a specific point of the wall, and h is the exchange coefficient, specifying the heat flow per unit area (h_x, h_y, h_z).
However, since the region where heat is generated is confined to the CF portion within the oxide, the NBCs actually resemble DBCs, with a fixed boundary temperature equal to the ambient one. This corresponds to the time-resolved equation:

$$\frac{\partial T}{\partial x} = \frac{T_1^{n+1} - T_{amb}}{dx} \tag{3.34}$$

Whose heat exchange parameter was ultimately defined by:

$$h = h_D = \frac{k_{TH}}{dx} \tag{3.35}$$

Note that as h decreases, the system approaches the adiabatic case, where no heat exchange occurs at the domain boundaries.
During the forming phase, temperature values larger than 1000K were registered in the central portion of the conducting filament.

3.2.8 System Update and Stop of the Simulation

While reaching the conclusion of the KMC iteration, several parameters were updated according to the outcomes of the simulated processes.
Transition rates were first extracted based on values such as local temperature and electric field. These rates were, then, used to compute the transition probabilities, with the model stochastically selecting the events that are going to happen, thereby updating the positions of the particles taking part in the switching mechanism.
Going on with the iterations, the global simulation time was incremented and other parameters, like time, current, and voltage were compared versus pre-determined targets[46]. This comparison was pivotal to determine whether a simulation should finish. If this is not the case, the next iteration starts and the process repeats.

Retention simulations

In the following Chapter, the potential of the KMC model is deeply investigated, with the final purpose of assess the retention characteristics of an OxRAM device.
After having defined the criteria adopted to study retention, first, the effect of the top Ti/HfO_2 interface is attentive. Next, an oxygen vacancy diffusion model is developed, where V_O appears in three distinct charge states: neutral, singly and doubly charged. Lastly, the activation energy configuration ensuring 10 years-retention is derived.

4.1 What is retention and how it is calculated

Leveraging on the stochastic property of the KMC model, the intrinsic device-to-device (D2D) variability of a memory array can be studied in terms of: probability distributions, Weibull curves and Time-to-Failure (TTF) plots.
First of all, it is essential to make a distinction between the meaning of Endurance and Retention. Even if both of them are exploited to characterize the reliability of an electronic device, having a higher endurance, most of the times, does not imply a high retention and vice-versa[35].
As depicted in Fig41a), endurance highlights the possibility of ensuring a change of state, either from LRS (ON) to HRS (OFF) or vice versa, for multiple writing/reading cycles. In particular, in Ref. **[37]**, high level of endurance[1] were proved up to 10^6 writing cycles, over which the LRS-HRS window margin was not ensured anymore.
On the other hand, retention gives a measure of the memory cell's ability to hold the data stored for a certain period of time without manifesting any decay. Fig41b) shows the retention characteristic of a 256kb array annealed at 150°C for increasing baking time, up to 1000h[58]. Overall, the two current values were highly retained over time, with a progressive LRS resistance degradation due to a small cross section of the CF[58].
In the successive simulation results, it will shown that this retention degradation is highly related to the oxygen vacancies concentration and their respective locations within the active medium.

[1]the OxRAM can endure more than 4σ distributions at 10^5 cycles

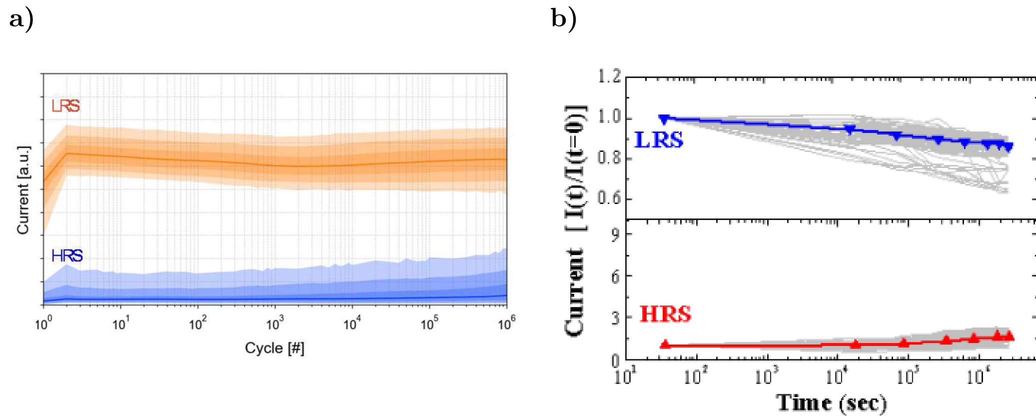


Figure 41: a) Endurance expressed in terms of single-cell current evolution with cycling, for HRS and LRS. Solid lines represent the median value and 1σ 2σ 3σ contours are showed in light colours. Adapted from **[37]**. b) Retention expressed in terms of normalized current evolution up to 1000h at 150°C annealing for 256-kb 1T1R array. Adapted from **[58]**.

To predict the device's reliability over multiple cycles and extended periods, of the order of years, it is not practical to merely wait and monitor how data is maintained. Therefore, an alternative faster method must be utilized.

The technique commonly employed among the R&D community is called *Accelerated Life Testing* (ALT). As implied by the name, it consists in several annealing tests conducted inside high-temperature ovens for various, more practical, baking times. Once the annealing is completed, the response of the memory to the applied thermal stress is analyzed to determine whether retention stability is ensured[59]. In fact, an elevated thermal energy supply may enhance the degradation process inside the material stack, leading to filament dissolution.

Taking as reference an LRS-programmed OxRAM cell, there are mainly two possible scenarios affecting resistive switching.[37][58][59]

- Thermal-induced oxygen vacancy diffusion within the resistive layer, according to the V_O charge state and locations (as described by first-principle theories of Par.2.4.3);
- Interface reactions, involving the recombination of O^{2-} ions, emitted from the top Ti layer, with V_O^{2+} within the dielectric. Consequently, the overall defect concentration diminishes, a non-negligible number of voids originates in the filament, and the total current flow is reduced.

Actually, thermal stress can cause structural alterations within the HfO_2 layer, too, such as grain growth or phase transitions[60]. However, in the present work, monoclinic Hafnia is assumed everywhere.

Paragraph 4.2.1 thoroughly explores the potential phenomena within an OxRAM cell during retention testing, emphasizing the role of activation energy in retention enhancement.

4.1.1 Aim of the simulations

The objective of this work is try to give a physical explanation to the retention experimental results reported in Fig.42 and in literature **[37]**. In *Weebit Nano*[37] it has been found that information was well retained after 10^4 cycling operations and ALT at 210°C up to 15h baking (Fig.42a). Whereas from the Arrhenius extrapolation of Fig.42b) 10 years retention up to 125°C[40] were predicted, with negligible impact of cycling on the final TTF line slope.

The findings presented in the next Sections aim to bridge the gap between simulation and experimental data, specifically addressing the slope value of 1.5eV.

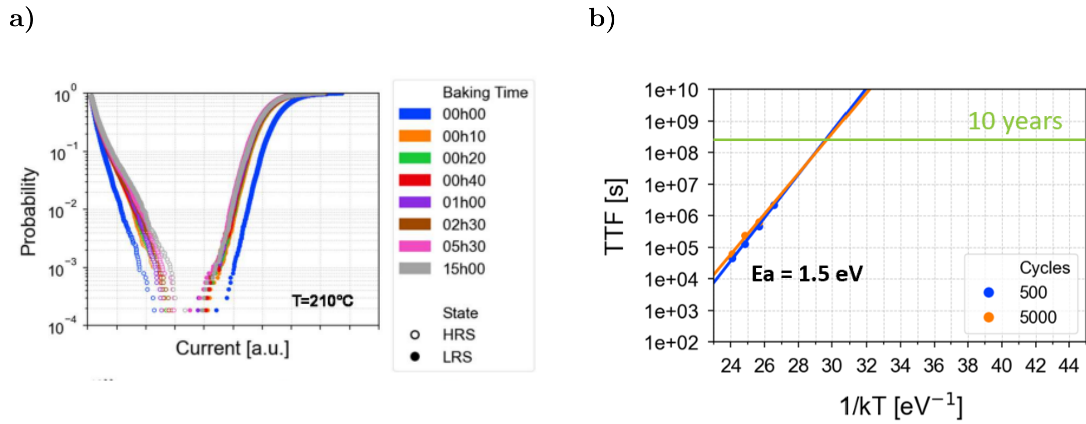


Figure 42: a) Experimental LRS and HRS current distributions for memory cells at 210°C for increasing baking time. Resistance degradation can be observed at the tails. Adapted from **[37]** b) Impact of cycling on TTF.

In terms of 10-years maximum temperature, typical commercial NVM applications require to store data at least at 85°C[62]. However, the ultimate target is to meet the requirements of the Automotive Electronics Council (AEC), whose applications can be classified with a different grade, depending on the temperature range employed [61].

AEC Grade	Operating Temperature
0	-40 to +150°C
1	-40 to +125°C
2	-40 to +105°C
3	-40 to +85°C

Table 4: AEC High Temperature Storage Requirements. Adapted from **[61]**

4.1.2 Bit error rate (BER) and failing bits

Though OxRAM cells has achieved promising reliability results in recent years[37], the stochastic nature of the failure mechanism still must be fully understood to guarantee a low Bit Error Rate (BER) in commercial memory applications[62].
In this study, as seen from the current distribution of Fig.42a), a progressive spread originating from the tails can be observed for increasing baking time. This effect is often attributed to some failing bits within the memory cells, which could pose a significant limitation to either a single device or to the whole memory architecture[61][62]. The higher is the time (and temperature), the closer will be the margin window between LRS and HRS, leading, finally, to data loss.
The tail spreading is clearly observable in the R_{LRS} CDF shown in Fig.43a), and the impact of time and temperature is validated by the Weibull plots in Fig.43b).
Both CDF and Weibull extraction will be pivotal in tackling single-cell data failure of retention.

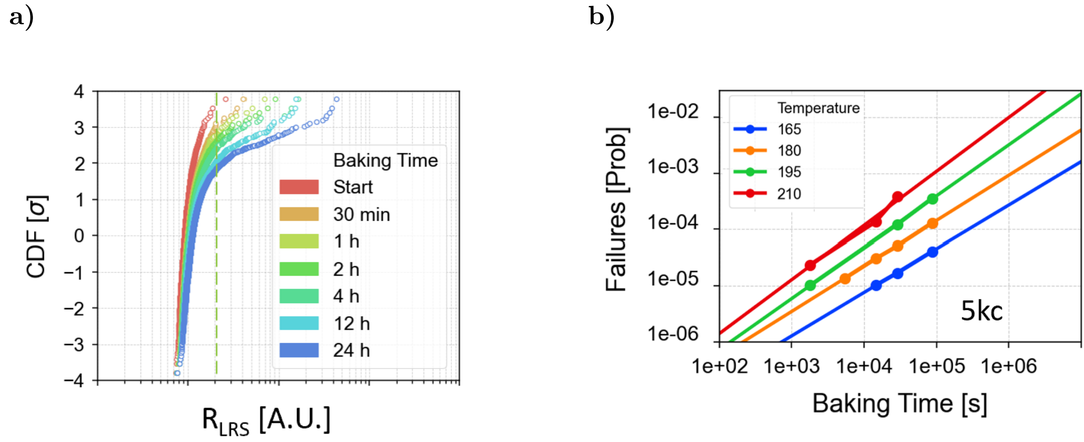


Figure 43: a) Experimental LRS resistance distribution for increasing baking time. Resistance degradation can be observed at very low failing probabilities. b) Example of Weibull plot for different temperatures and baking times.

According to literature [**49**][**62**] the failing bits phenomenon could be addressed to early (ms range) relaxation effects, occurring just after programming and due to fast redistribution of oxygen vacancies and unwanted generation/recombination.
Nevertheless, this work mainly focuses on retention analysis, accounting for longer times, with relaxation effects deemed negligible.

4.2 Initial Setup: 100 OxRAM LRS programmed

Simulation data were extracted starting from 100 ReRAM cells, on which Forming operation was already applied, according to the scheme described in Sec.3.2.
Forming is a crucial step for determining the final reliability of the device, since the right combination of setup parameters must be selected, in order to achieve a robust filament. In the following, the user input parameters chosen for these specific tests are reported, along with a snapshot of cell number 0.

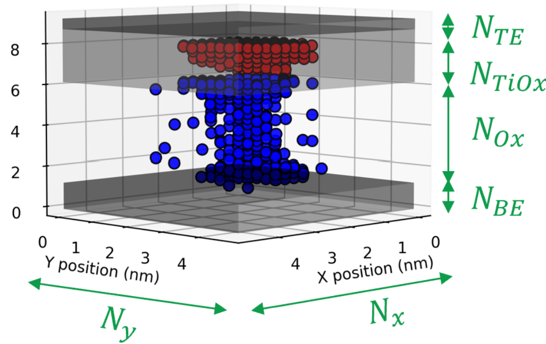


Figure 44: Final snapshot of cell #0 just after forming. Its resistance is $R_{LRS} = 14.88k\Omega$

Parameter	value	Parameter	value
d	2.6 $\mathring{A}$	I_{CC}	100μA
$N_x = N_y = N_{ox}$	19	V_{max}	2V
N_{TE}	2	V_{step}	0.1V
N_{TiO_x}	10	$t_{ramp,up}$	100ns
N_{BE}	5	$E_{A,G}$	2.8eV
ε_{r,HfO_2}	16	$E_{A,R}$	0.2eV
ε_{r,TiO_x}	1024	E_{A,V_0}	1.4eV
$nmos\ carac$	None	b	0.4

Table 5: Setup parameters employed for the forming of 100 OxRAM cells

No transistor compliance was applied. The programming current was set to the maximum one considering a sort of current mirror source of $100\mu A$.
The methods most frequently employed for the characterization of the statistical fluctuations of the tests were:

- Probability Density Function (PDF);
- Cumulative Density Function (CDF), typically a log-normal distribution;
- Quantile-Quantile (QQ) Plot, where a straighter line indicates that the density curve is more similar to a Gaussian normal distribution.

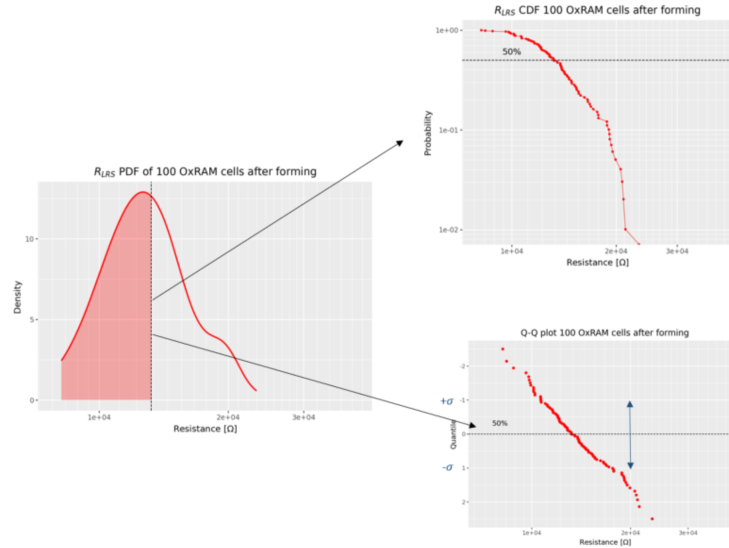


Figure 45: Based on the 100 OxRAM R_{LRS} value read at $V_{read} = 0.1V$, the PDF is extracted. From this, CDF distribution and QQ plot are derived. The 50% value of the CDF correspond to the median, equal to $R_{LRS} = 12320.4\Omega$

Once having defined the starting point, several LRS retention simulations were carried out for different baking temperatures, ranging from 100°C up to 230°C, and baking times, whose values spanned from few minutes up to 10 hours.

4.2.1 Activation energies role

As it will thoroughly discussed, talking about temperature and time is not enough to address retention behavior, since it strongly depends on the the microstructural properties of the MIM stack[37].
Within the KMC framework, indeed, every layer must be properly engineered in terms of its material features to mitigate resistance variability over time.
In this context, such features are the activation energies for V_O O^{2-} generation, recombination and respective diffusion, whose values change according to the material.

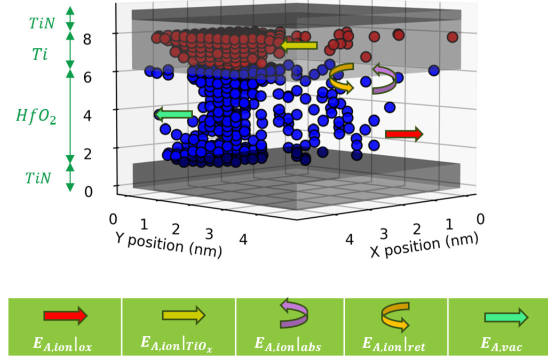


Figure 46: Sketch of the main activation energies ruling retention on LRS OxRAM cell #20 after forming.

In Fig.46 it is reported a sketch of the main diffusion mechanisms taking place throughout the MIM stack. Indeed, as already discussed in Par.3.1.2, the diffusion of a charge specie in OxRAM devices is intrinsically linked to its activation energy. The latter is defined as the minimum energy required to transit from one stable position to another within the material's lattice. The correlation is well explained by the Arrhenius law (eq.3.2), which states that the diffusion rate exponentially decreases for higher activation energy.
In particular:

- Regarding oxygen vacancies ($E_{A,vac}$), homogeneous (isotropic) diffusion was taken into account and it was possible only inside the resistive medium;
- On the contrary, oxygen ions were free to move both within the oxide ($E_{A,ion}|_{ox}$) and the scavenging layer ($E_{A,ion}|_{TiO_x}$), since at the top interface a non-stoichiometric film of TiO_x was formed. The activation energy for ion migration inside Hafnia was typically lower than that within Ti, because of its stronger reactivity with oxygen;
- In conjunction with these, between 6.24 and 6.5 nm[2], two more activation energies were defined. These represent the emission ($E_{A,ion}|_{em}$) and absorption barrier ($E_{A,ion}|_{abs}$) energy from/to Ti to/from HfO_2, respectively. This implies that the interface has to be treated as an additional layer comprising the stack, and it has a non-negligible role in the resistance drift characterization.

In conclusion, low E_A would require fewer thermal excitation to make the particle overcoming the energy barrier, leading to fast switching, especially under thermal stress. Conversely, higher E_A might result in more stable data, but require higher voltages or longer times during switching.

[2]namely, one step mesh above the interface, fixed at 6.24nm

4.3 Simulation framework

The simulations discussed in the subsequent Sections can be categorized into two classes, based on the object of study (Fig.47).

1. A first attempt was to focus solely on the role of oxygen ions diffusion, and their recombination with oxygen vacancies at the top Ti/HfO_2 interface. The reason behind this choice was to understand the role of the sole interface in ReRAM stabilty especially during retention experiments, since most of the times the rupture of the CF originates from this point, creating a gap that, at the end, would drift the resistance to HRS values[64].
   Therefore, following the layout of Fig.46, the activation energies for ion diffusion through the resistive and scavenging layers, along with the absorption and emission barriers at the interface, were adjusted individually to observe their impacts and to determine which one might be the dominant mechanism affecting CF stability.
   Among the several energy configurations explored, a final one was selected. Using this configuration, the device's lifetime was estimated by employing Weibull and TTF graphs.

2. Once having assessed this first section, more consistent results were extracted when considering also the role of oxygen vacancy for filament dissolution.[9][10][48]
   To this purpose, as previously mentioned in Par.2.4.3, three V_O species can coexist within the resistive medium. This applies to both scenarios: either with three different charge states or solely with neutral V_O, following filament stabilization.

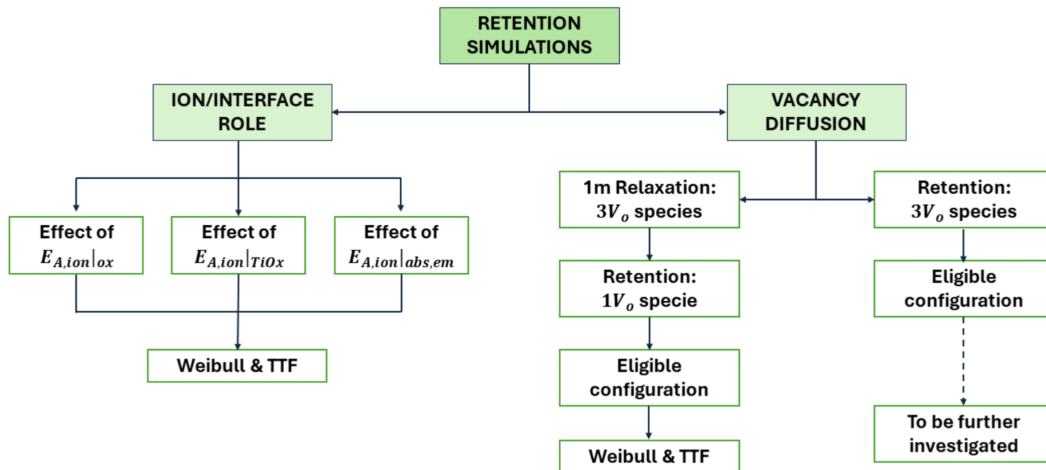


Figure 47: Simulation framework

4.4 Ion-Interface Role

4.4.1 Effect of Temperature

Before diving into the core of retention simulations, some preliminary tests were carried out for 1 hour baking at different temperatures. The objective is to see how thermal energy is able to modify the shape of the filament and affect final conduction.

In line with the approach of *Perez* [**59**], simulations were conducted at both low and high temperatures. This methodology allows for simulating real experimental low temperatures while also pushing forward the retention within a short period, thanks to higher T.

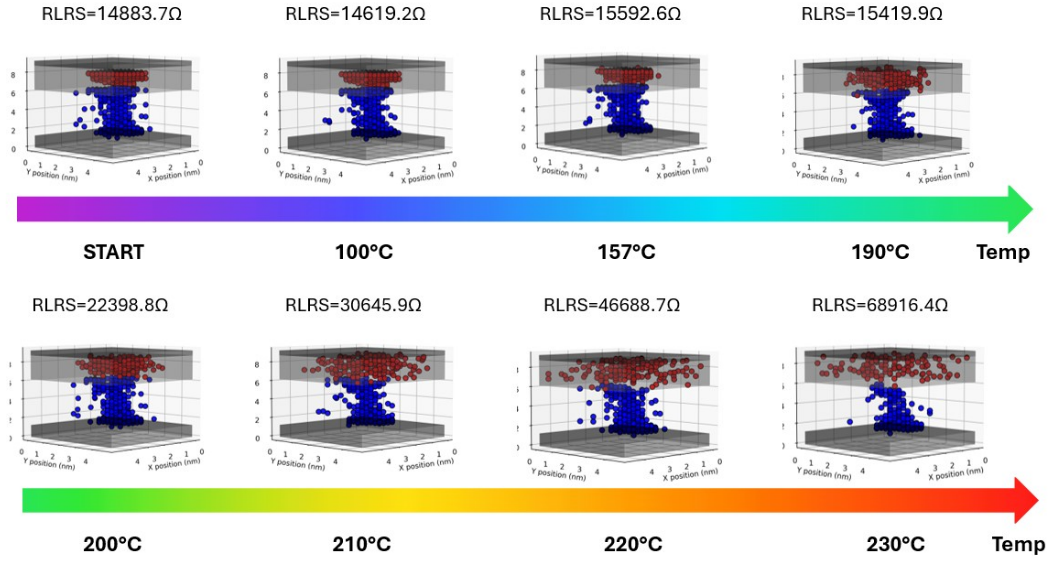


Figure 48: Effect of temperature for 1h baking on cell #0. An increase of resistance is registered, leading to a less stable filament.

As expected, the higher is temperature, the more particles gain kinetic energy, and the more drift of resistance is displayed in Fig48, maintaining an order of magnitude of $10^4\Omega$. In this particular case, since only interface effects were considered, the rupture of the filament starts from the top, where an increasing diffusion of ions is registered. The last statement is confirmed by the insights of Fig.49, regarding only three temperatures among the seven presented.
Regarding the simulation setup, the default parameters are listed in Table6.

$E_{A,ion}\vert_{ox}$	$E_{A,ion}\vert_{TiOx}$	$E_{A,ion}\vert_{abs}$	$E_{A,ion}\vert_{em}$
0.6eV	1.6eV	0.4eV	1.4eV

Table 6: Activation energies setup

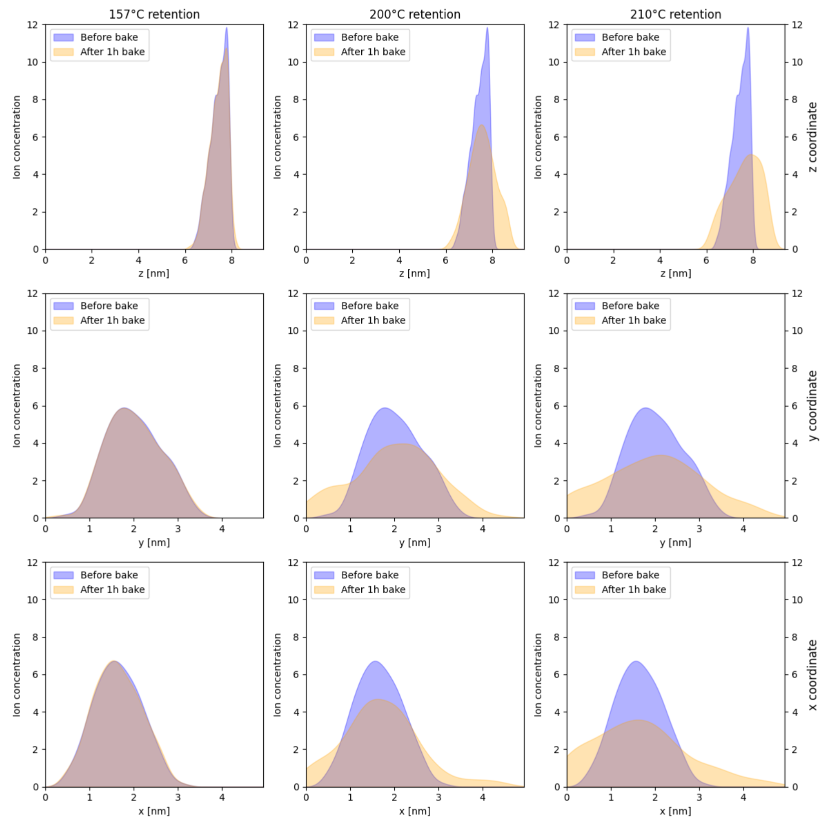


Figure 49: Oxygen ions density distribution along x, y, z coordinates before (blue) and after (orange) 1h bake at 157°C, 200°C, 210°C.

The stability of a particular simulation condition was determined in terms of the statistical methodology mentioned in Section 4.2 (Fig.45), but another instrument could be by looking at the way vacancies diffuse out of the filament (50). The more distant they are, the more voids are present in between the percolation paths, and the less conductive will result to be the memory cell.

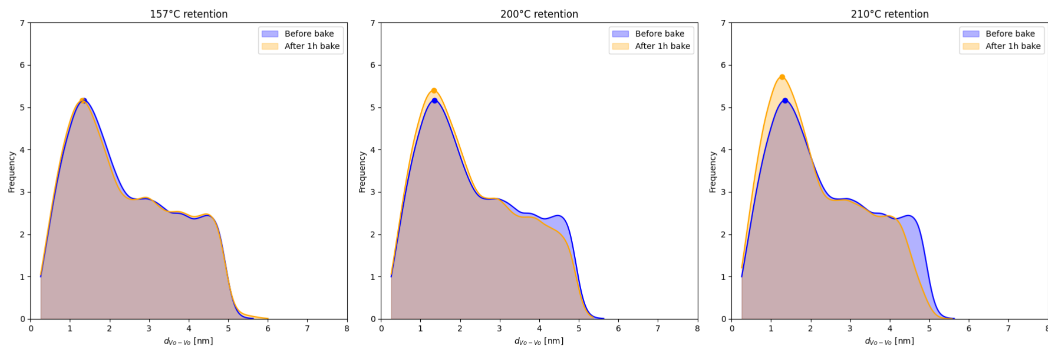


Figure 50: Simulated distribution of the distance between adjacent oxygen vacancies within the resistive layer, before and after 1h bake at 157°C, 200°C, 210°C

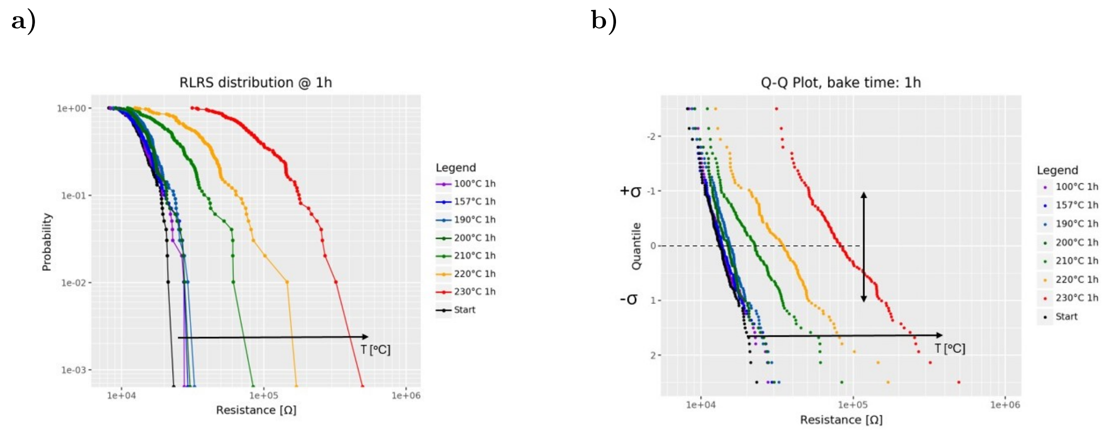


Figure 51: Resistance variability and drift for increasing baking temperature and fixed time (1h). a) LRS distribution b) QQ-plot.

Observe that with rising temperature, there is a noticeable rightward shift in the distribution. In contrast, Fig.43a) illustrates that resistance drift is mainly present at the distribution tails. This discrepancy may arise because the diffusion dynamics of oxygen vacancies, influenced by crystal anisotropy[10], are not accounted for, and also, other possible failure/diffusion mechanisms may not have been taken into account.

4.4.2 Effect of Baking Time

Opposite to what was done up to now, in order to characterize the effect of baking time, the temperature was maintained fixed to a value of 210°C. Then, five times were investigated, from 6 minutes up to 10 hour baking.
The same simulation setup from table 6, was chosen.

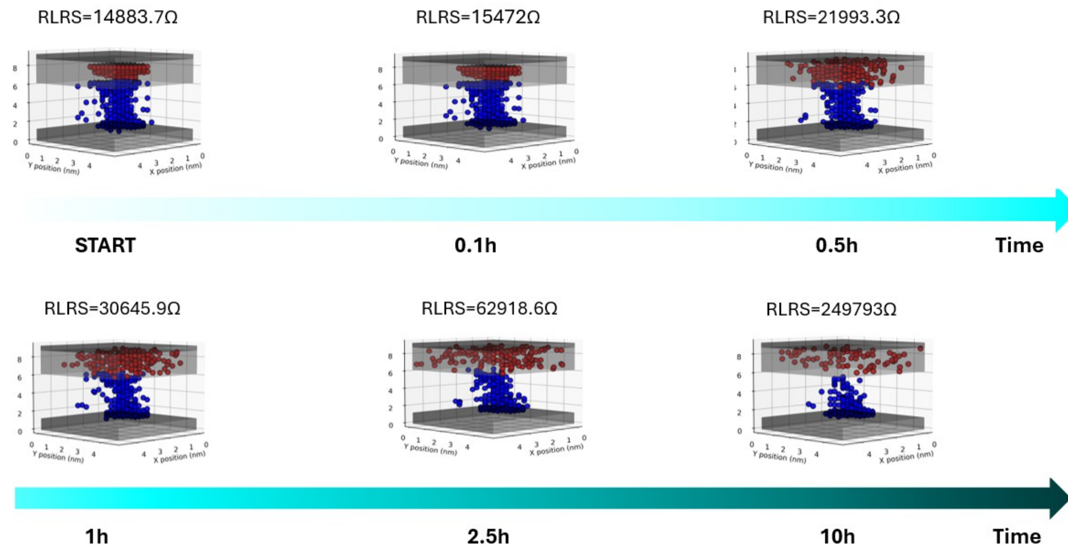


Figure 52: Snapshots of cell #0 for increasing baking time at 210°C. CF breaking is impacted from the top interface.

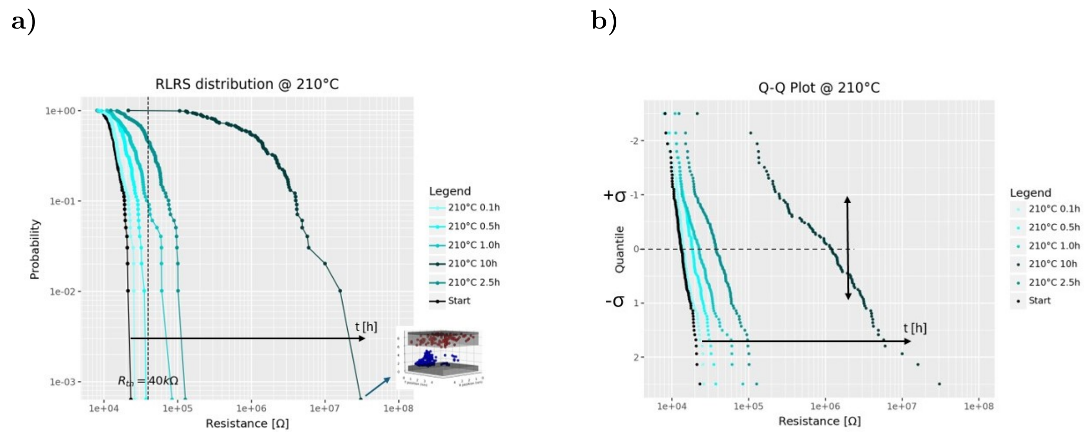


Figure 53: Resistance variability and drift for increasing baking time and fixed temperature (210°C). a) LRS distribution b) QQ-plot.

Baking time, as well as thermal energy, had the effect of enhancing filament dissolution and its consecutive rupture from the top interface.
When considering a resistance threshold value of $R_{th} = 40k\Omega$ for the low-resistive state, it is immediately clear from the distribution of Fig 53a) that the 99% of the cells fail for the 10 hours baking. Hence, in order to finely tune the cell-to-cell variability, maximum times of 2.5 hours were taken into account for the next simulations.
At the same time, selecting a lower temperature permits the use of longer duration.

4.4.3 Ion diffusion within the resistive layer

This section revolves around the investigation of ion diffusion throughout the memory cell.
The first step to examine this effect involves altering the activation energy for ion migration within the oxide by $\pm 200meV$ with respect to the default value of Table 6.
Three energy setup are going to be discussed.

$E_{A,ion}\|_{ox}$	$E_{A,ion}\|_{TiOx}$	$E_{A,ion}\|_{abs}$	$E_{A,ion}\|_{em}$
0.4eV	1.6eV	0.4eV	1.4eV
0.6eV	1.6eV	0.4eV	1.4eV
0.8eV	1.6eV	0.4eV	1.4eV

Table 7: Energy setup for varying $E_{A,ion}|_{ox}$

Initially, ions were not present in the oxide, they just accumulated within the scavenging TiO_x layer, once forming operation was completed. In order to diffuse inside Hafnia, they, first, had to cross the emission energy barrier, which was equal to 1.4eV for all the three configurations. Once did so, the ion could either recombine with a neighbouring vacancy, migrate inside the resistive layer or being absorbed back. The low activation energy for recombination ($E_{A,R} = 0.2eV$), suggests that recombination was the most feasible option, because according to the theory of Chapter3, to it corresponds the lowest rate and, thus, the highest event probability.

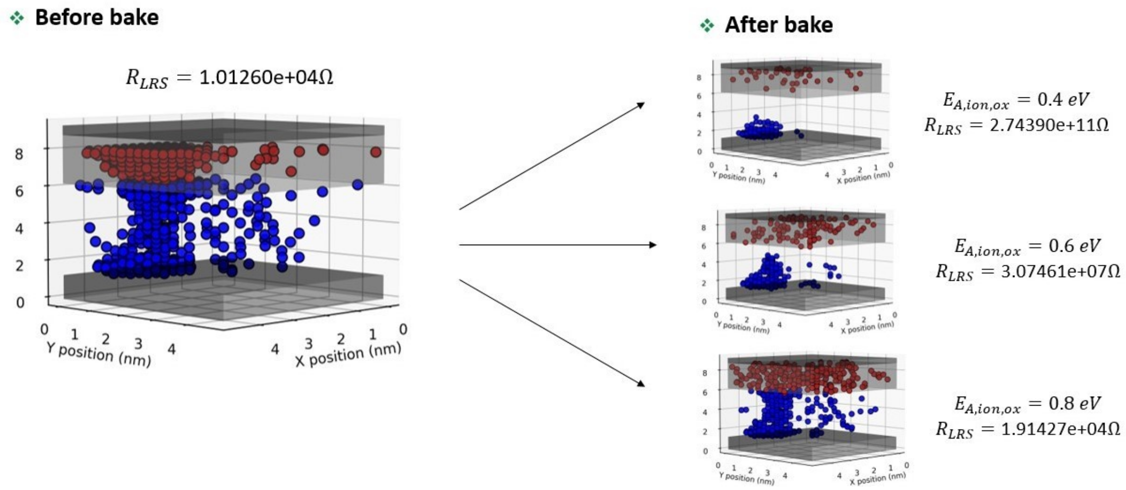


Figure 54: Effect of variable $E_{A,ion}|_{ox}$ on cell #20 for 210°C 10h retention simulation.

By observing the snapshots in Fig.54, only the 0.8eV configuration cell showed a resistance value below the threshold after 10 hours baking. This indicates that a more stable filament was achieved for $E_{A,ion}|_{ox} = 0.8eV$, whereas lower activation energies led to ion depletion in the scavenging layer and additional recombination events. These assumptions, derived from single cell results, were validated by the subsequent distribution and QQ plot taken at 2.5 hours ALT.

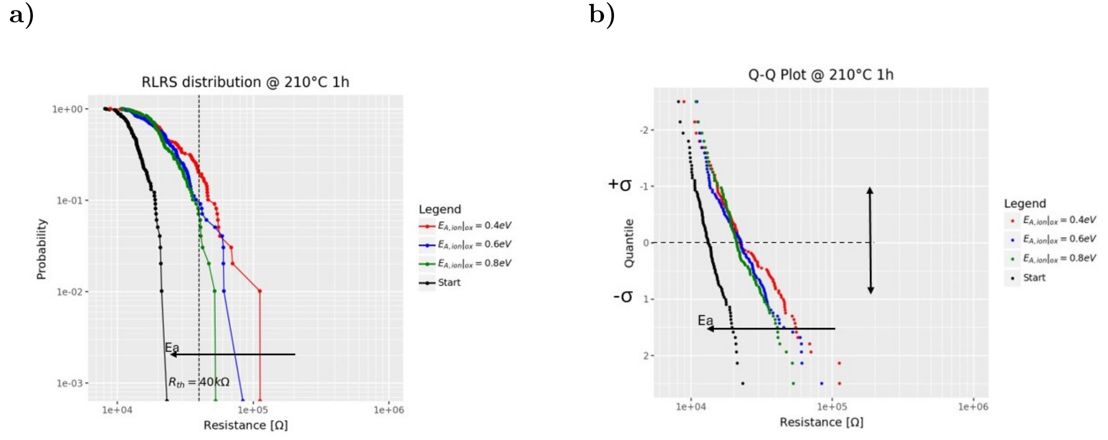


Figure 55: Resistance variability and drift for variable $E_{A,ion}|_{ox}$ at 210°C 1h. a) LRS distribution b) QQ-plot.

The same simulation results can be translated in terms of mean resistance for varying baking time, up to 3 hours. Following a 3rd order polynomial, in Fig.56, the coefficients a, b, c, d were extracted and the behavior for $E_{A,ion}|_{ox} = 0.1eV,\ 0.2eV$ were derived without the need of further calculations.

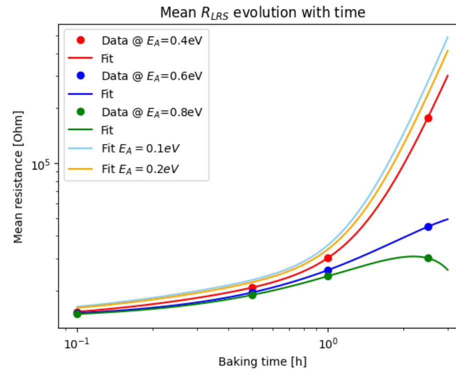


Figure 56: Mean resistance evolution with time for varying $E_{A,ion}|_{ox}$

4.4.4 Ion diffusion within the scavenging layer

The same procedure was applied to the activation energy responsible for ion migration inside TiO_x, and whose default value was varied by $\pm 200meV$.
The simulation configuration are listed in the following.

$E_{A,ion}\|_{ox}$	$E_{A,ion}\|_{TiOx}$	$E_{A,ion}\|_{abs}$	$E_{A,ion}\|_{em}$
0.6eV	1.4eV	0.4eV	1.4eV
0.6eV	1.6eV	0.4eV	1.4eV
0.6eV	1.8eV	0.4eV	1.4eV

Table 8: Energy setup for varying $E_{A,ion}|_{TiOx}$

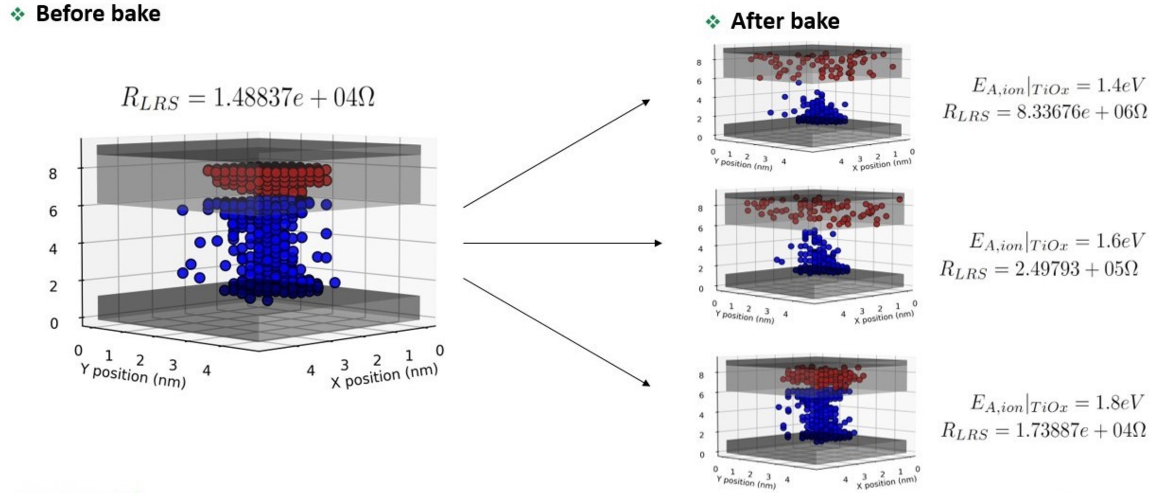


Figure 57: Effect of variable $E_{A,ion}|_{TiOx}$ on cell #0 for 210°C 10h retention simulation.

In accordance to what was illustrated previously, very stable results were obtained for larger energies (Fig.57).
In terms of LRS distribution, the 1.8eV case exhibited minimal variations with respect to the starting condition, as confirmed by Fig.58. In conjunction with this, computational time was more affordable too.

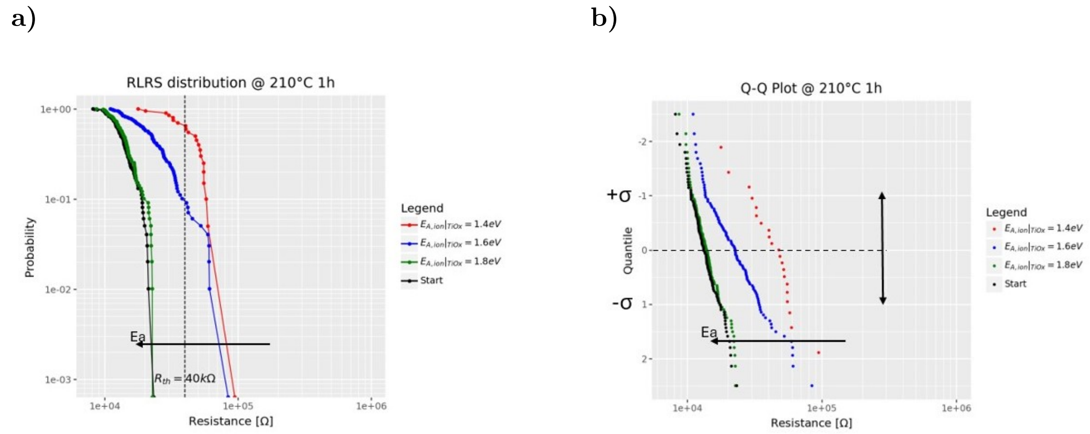


Figure 58: Resistance variability and drift for variable $E_{A,ion}|_{TiOx}$ at 210°C 1h. a) LRS distribution b) QQ-plot.

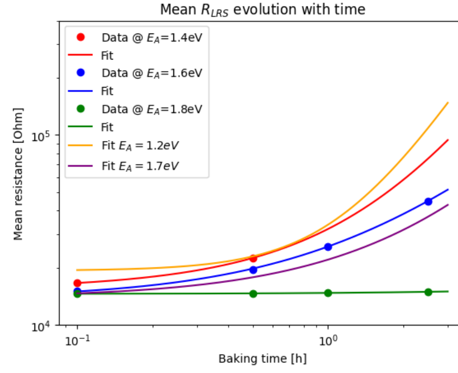


Figure 59: Mean resistance evolution with time for varying $E_{A,ion}|_{TiOx}$

The more stable LRS achieved for 1.8eV demonstrate that the final resistance remained almost constant, even over extended periods. This was because ions were less likely to migrate within the substoichiometric layer or be released from it under such conditions.

4.4.5 Effect of absorption and emission barriers

Ultimately, the effect of absorption and emission barriers at the Ti/HfO_2 interface was investigated. To this purpose, both emission and absorption energy were increased of $+200meV$ with respect to the standard of Table6. In this way, a sort of continuum between ion diffusion within Ti and HfO_2 was assumed, thus neglecting any variation in correspondence of the interface.

$E_{A,ion}\|_{ox}$	$E_{A,ion}\|_{TiOx}$	$E_{A,ion}\|_{abs}$	$E_{A,ion}\|_{em}$
0.6eV	1.6eV	0.4eV	1.4eV
0.6eV	1.6eV	0.6eV	1.6eV

Table 9: Energy setup for varying $E_{A,ion}|_{abs}$ and $E_{A,ion}|_{em}$

CDF and QQ-plot of the 100 cells distribution after 1 hour 210°C baking, are reported in Fig.60

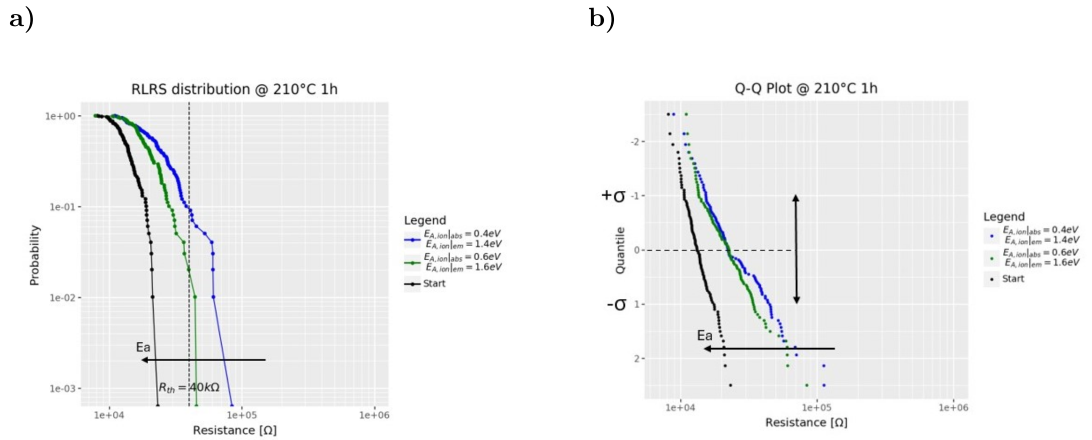


Figure 60: Resistance variability and drift for variable $E_{A,ion}|_{abs}$ and $E_{A,ion}|_{em}$ at 210°C 1h. a) LRS distribution b) QQ-plot.

These last plots are consistent with what was asserted up to now: the lower is the activation

energy, the more the stability of the bit is at risk.

4.4.6 Failure probability

Despite the best results were achieved for the configuration at $E_{A,ion}|_{TiOx} = 1.8eV$, with a R_{LRS} distribution almost overlapping to the starting one, the default configuration(Table 6) was chosen as the elective one to be further analyzed.
This paragraph aims to reinterpret the RS dynamics using the *Failure Probability* concept to evaluate the long-term reliability of the setup.
First of all, a threshold value for the resistance must be set. This implies that for $R > R_{th}$, the LRS is not stable anymore and the information is assumed to be lost. In this work, the threshold was assumed to be 40kΩ, corresponding approximately to three times the mean R_{LRS} value of the 100 OxRAM cells before testing, equal to 13kΩ.
In the following an example of cell at 13kΩ and 40kΩ is shown.

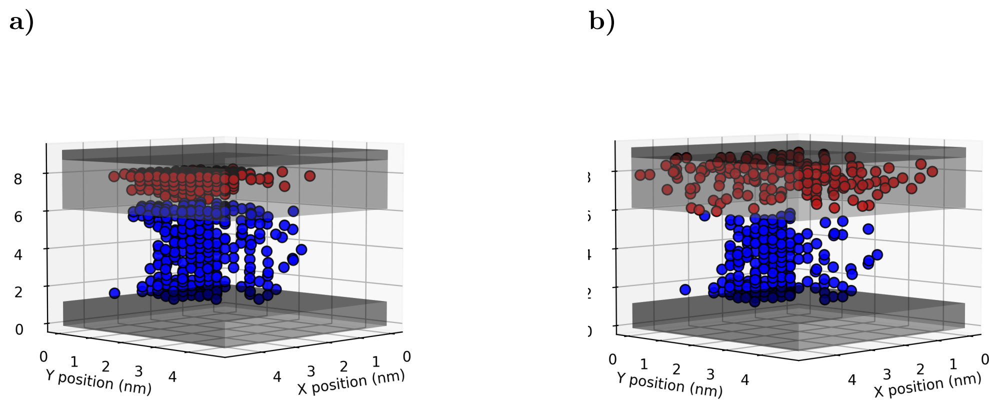


Figure 61: Cell #87 with a) $R = 13k\Omega$ and b) $R = 45k\Omega$.

In this particular case, CF breaking occured at the top interface, where the absence of point contacts made the conductivity to switch to the HRS.
Next, the procedure consisted in the individuation of the failure percentage corresponding to $R \equiv R_{th}$, starting from the fitting of the LRS distribution curves.

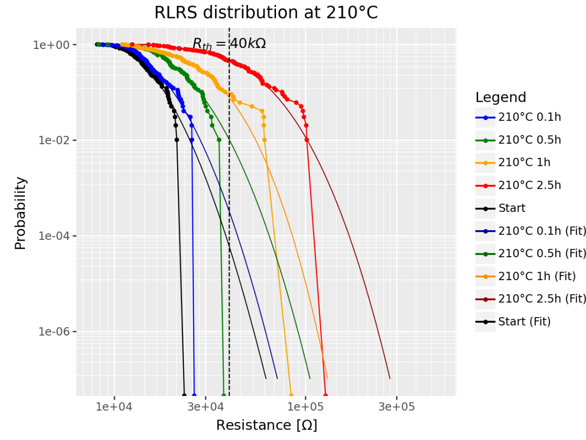


Figure 62: RLRS distribution at 210°C for varying baking time: simulation data and fitted curves. Fitting is performed neglecting the tails.

The corresponding failure probabilities, expressed in percentage value, are reported in the table below.

Bake time	Failure [%]
0.1h	0.083%
0.5h	1.2%
1.0h	7.73%
2.5h	46.5%

Table 10: Percentage of failure at 210°C, default energies, for increasing baking time. Failure probability increases along with time.

The same procedure was applied to other temperatures for a bunch of simulations: 190°C, 200°C, 220°C.
The extracted failing probabilities are resumed in table11.

Temperature[°C]	Bake time[h]	Failure [%]
190°C	0.1h	0.00021%
190°C	0.5h	0.00034%
190°C	1.0h	0.0042%
190°C	2.5h	0.0089%
200°C	0.1h	0.0087%
200°C	0.5h	0.032%
200°C	1.0h	0.178%
200°C	2.5h	5.534%
210°C	0.1h	0.083%
210°C	0.5h	1.2%
210°C	1.0h	7.73%
210°C	2.5h	46.5%
220°C	0.1h	8.24%
220°C	0.5h	35.49%
220°C	1.0h	68.87%
220°C	2.5h	100%

Table 11: Increasing percentage of failure at default energies, for increasing baking time and baking temperature.

4.4.7 Weibull graph

Using Weibull distribution, the trend of failure probability with time is described by[65]:

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right] \tag{4.1}$$

where α and β are the scale and shape parameter, respectively.
Once the failing data are collected for all the simulation setup, they must be linearized according to the formula[65]:

$$\ln[-\ln(1 - F(t))] = \beta\ln(t)\beta\ln(\alpha) \tag{4.2}$$

From this straightforward calculation, the Weibull plot was ready to be extracted.

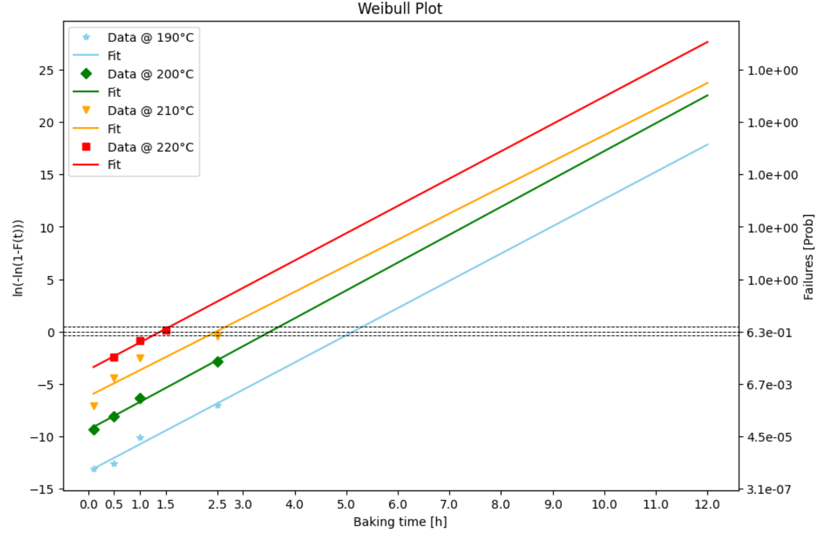


Figure 63: Weibull graph for default energy simulations extracted at 4 different temperatures and baking times. The three dotted lines correspond to 50%, 63.2% and 80% probability of failure.

Each data point corresponds to a specific simulation.
Referring to the graph, the α parameter is given by the x value for which the Weibull curve intersects the 63.2% failing probability line, and so the baking time for which a particular simulation had a 63.2% probability of failing.
Whereas, the β parameter corresponds to the slope of the Weibull line.
Both scale and shape parameters are resumed in the table below.

Temperature[°C]	α[h]	β
190°C	5.14 h	2.6
200°C	3.52 h	2.66
210°C	2.48 h	2.5
220°C	1.4 h	2.6

Table 12: Scale (α) and shape (β) parameter for the 4 different Weibull lines

To conclude, Weibull graph represents a practical instrument to look directly at the failing times distributions of the OxRAM cells, for varying Temperature.
Since the slopes of the line are almost similar, it can be derived that the occurring failing mechanism is similar, independently from Temperature. This property makes the behavior of the cells more predictable, allowing for better modeling their reliability.

4.4.8 Time to Failure

The last point of the retention and reliability analysis of the device consists in the Time-to-Failure plot. In this graph, also known with the term of *Arrhenius Plot*, the device lifetime was reported from the inverse formula of 4.2.

$$TTF = \alpha(-\ln(1-q))^{1/\beta} \tag{4.3}$$

Where α and β were given from Table 12 and q referred to a particular fail probability. In this work $q = F(t) = 0.5,\ 0.632,\ 0.8$.

By plotting $\ln(TTF)$ against $q/(k_B T)^3$, then, the Time-to-Failure graph was derived.

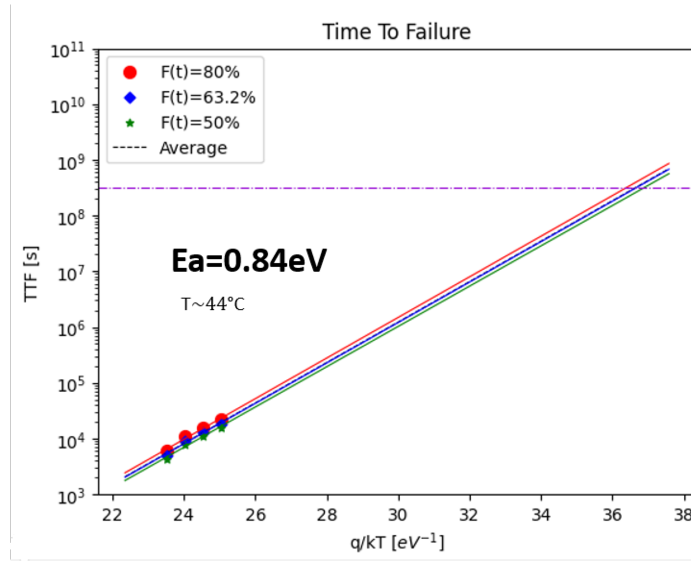


Figure 64: Time-to-Failure Plot of the default energy simulation setup. The 3 curves corresponds to three failing probabilities, respectively: 50% (green), 63.2% (blue) and 80% (red). The dotted black line is the average of the three. The 4 points, from right to left, represents the 4 temperatures: 190°C, 200°C, 210°C, 220°C. The slope of the average line is $E_A = 0.84eV$ and the maximum 10ys temperature is 44°C.

The term "Arrhenius" came from the principle that the temperature dependence of failure mechanisms follows the Arrhenius law, which allows TTF to be represented as:

$$TTF = A\exp \cdot \left(\frac{E_A}{k_B T}\right) \tag{4.4}$$

where A is a pre-exponential factor, and E_A is the activation energy ruling the failing process during retention tests.
Being:

$$\ln(TTF) = \ln A + \frac{E_A}{k_B T} \tag{4.5}$$

referring to Fig.64, E_A corresponds to the slope of the line, equal to 0.84eV.
In accordance to the experimental TTF of Fig.42b), the higher is the activation energy, the steeper is the TTF line and the higher is the maximum temperature at which the device can work for a time duration of 10 years.
In this particular case, the temperature was derived from the intersection of the average TTF with the 10ys line, and it corresponded to:

$$\frac{q}{k_B T} = 36.6eV^{-1} \quad \longrightarrow \quad T = 317K \quad \longrightarrow \quad T = 44°C$$

The extracted temperature value was far distant from the automotive goal of 150°C, indicative of the fact that this type of OxRAM cell, when in such operating conditions, did not give reliable results in the long term.
The lifetimes, the corresponding E_A and maximum T are reported for the three fail probabilities.

[3] q is the elementary charge, k_B is the Boltzmann constant and T is the temperature expressed in Kelvin

	$TTF_{190°C}$	$TTF_{200°C}$	$TTF_{210°C}$	$TTF_{220°C}$	E_A [eV]	T [°C]
50%	1.6×10^4s	1.1×10^4s	7.7×10^3s	4.38×10^3s	0.84eV	46°C
63.2%	1.85×10^4s	1.27×10^4s	8.9×10^3s	5.04×10^3s	0.84eV	43°C
80%	2.2×10^4s	1.52×10^4s	1.08×10^4s	6.05×10^3s	0.84eV	42°C

Table 13: Lifetimes, activation energy and maximum 10ys temperature extracted for three failing probabilities: 50%, 63.2%, 80%

Referring always to the energy setup of Table6, the origin of a 0.84eV slope could be related either to the ion diffusion energy within the resistive layer (=0.6eV) or to the ion emission barrier from Ti to HfO_2.
To understand which is the actual factor affecting the stability of the device under study, $E_{A,ion}|_{ox}$ was increased to 0.8eV.
The new energy parameters were:

$E_{A,ion}\|_{ox}$	$E_{A,ion}\|_{TiOx}$	$E_{A,ion}\|_{abs}$	$E_{A,ion}\|_{em}$
0.8eV	1.6eV	0.4eV	1.4eV

Table 14: Activation energies setup

The resulting Weibull and TTF graphs are outlined.

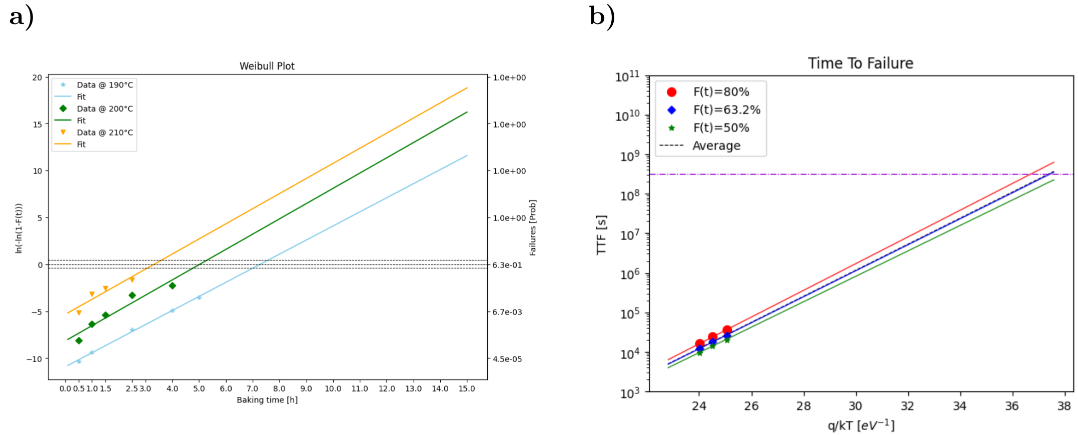


Figure 65: a) Weibull plot extracted for three different temperatures: 190°C (skyblue), 200°C (green), 210°C (orange). b) TTF graph for varying fail probability: 50% (green), 63.2% (blue), 80% (red). The final slope is 0.76eV and the max 10ys Temperature is 37°C.

Also in this case, the Weibull lines are almost parallel, with $\beta \approx 1.6$. On the other hand the slope of the TTF is $E_A = 0.76eV$, thus very similar to the previous one, confirming that the limiting mechanism could be the ion diffusion within the active medium. In order to enhance the reliability of the device, such migration energy should be further increased, in conjunction with the absorption barrier (=0.4eV), which probably keeps low the overall TTF slope.
In conclusion, the results presented up to this point did not fulfill the automotive requirement. This can be attributed to a poor characterization of the activation energies and, simultaneously, to the focus solely on ion interface effects, by ignoring vacancy diffusion.

4.5 Vacancy diffusion

The second part of the study aims to do an in-depth analysis of the physical mechanisms involved during retention tests. To this purpose, the characterization of the barrier energy for vacancy diffusion was modified throughout the code.
In Section 2.4.3, three different species of vacancies were introduced: neutral V_O^0, singly charged V_O^{1+} and doubly charged V_O^{2+}. Each of them has a distinct impact on the filament shape.
The way to discriminate one V_O specie from the other was by means of the neighbouring traps. In particular, the code implementation was based on the results from *Duncan et al.*, and assumed that:

- An isolated trap is doubly charged,
- A vacancy with one neighboring trap is singly charged,
- A cluster of vacancies is neutral.

The distinction was made exploiting the vacancy coordination number (*cv_oxvac*), which represents the number of neighboring vacancies and which is updated every KMC iteration. The classification followed the rules:

$$cv_oxvac = 0 \longrightarrow V_O^{2+}$$
$$cv_oxvac = 1 \longrightarrow V_O^{1+}$$
$$cv_oxvac \geq 2 \longrightarrow V_O^{0}$$

The activation energies were adjusted based on the charge state, according to the scheme:

$$V_O^{2+} \longrightarrow E_{A,vac}$$
$$V_O^{1+} \longrightarrow E_{A,vac} + 0.1eV$$
$$V_O^{0} \longrightarrow E_{A,vac} + 1eV$$

These considerations might account for different mobility and impact of each vacancy species on the memory cell stability, therefore providing a more accurate simulation of retention dynamics.
It is important to remind that diffusion is considered isotropic, thus equal for all the directions, independently from crystal orientation.
Referring to the simulation framework of Fig.47, two different approaches were studied for the characterization of filament dissolution as a result of oxygen vacancies diffusion.

1. Two-step like process. Here, a first 1 minute retention test was performed to mimic the relaxation occurring during the time interval between Forming operation and Retention testing. During this phase, the three aforementioned V_O species were taken into account.
   Successively, the retention simulations were carried out for larger baking times and several temperatures, considering only the dominating role of neutral traps. This is an underestimation of the real process but, according to Ref.**[66]**, since no bias was applied during ALTs, traps did not exhibit an ionization nature.

2. A more comprehensive study of the retention behavior when all three V_O charge states were involved.

Out of the two options, the first was favored, and the reason behind this preference will be thoroughly discussed in the upcoming sections.

4.6 Two-step retention simulation

4.6.1 First step: 1 minute relaxation

When calling the function which involves the characterization of three distinct vacancy species, a sort of "coalescence" of the filament was observed, as in Fig.66. This behavior was addressed to the fact that, in LRS, most of the traps were neutral, since they were close to the conductive filament region. Furthermore, due to the higher activation energy, the mobility of V_O^0 was greatly limited, favouring the association behavior close to the CF, as seen in Section2.4.3.

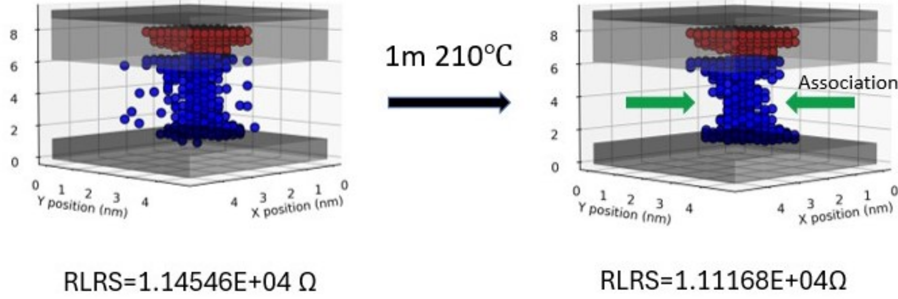


Figure 66: Snapshot of cell#0 before and after 1 minute retention at 210°C. The lower resistance is addressed to a thinning of the filament in the central portion.

Such simulation was carried out fixing all the energy parameters to 10eV, except for vacancies, whose diffusion energies are listed in the following:

$$V_O^{2+} \longrightarrow E_{A,vac} = 1eV$$
$$V_O^{1+} \longrightarrow E_{A,vac} = 1.1eV$$
$$V_O^{0} \longrightarrow E_{A,vac} = 2eV$$

In terms of resistance drift, no important variations were observed, as confirmed by the distribution of Fig.67. The aim of this first step retention was just to stabilize the filament.

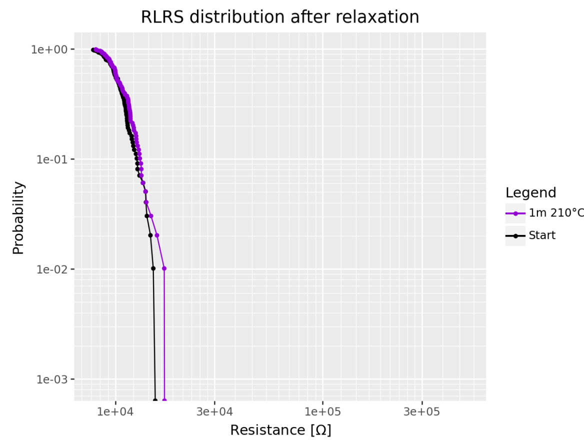


Figure 67: RLRS distribution after 1m 210°C baking, mimicking relaxation.

4.6.2 Second step: Retention

Once this was done, *accelerated-life-testing* could be carried out using the newly "relaxed" 100 OxRAM. The procedure was identical to the one described in Sec.4.4, with the exception that a real case study was examined. This necessitated a new energy configuration setup based on *ab initio*-related parameters.

Eligible configuration

Before reaching the optimal energy configuration, several tests were carried out, and the excursus is laid out in the upcoming lines.
The starting energy values are resumed in Table15.

$E_{A,rec}$	$E_{A,ion}\|_{ox}$	$E_{A,ion}\|_{TiOx}$	$E_{A,ion}\|_{abs}$	$E_{A,ion}\|_{em}$	$E_{A,vac}$
0.2eV	0.2eV	1.5eV	1.1eV	1.5eV	1.5eV

Table 15: Ab initio-related activation energies

Note that the vacancy activation energy $E_{A,vac}$ was related just to a single specie of V_O, since just neutral traps were assumed to be present within the active medium.
A first simulation was run for such setup for one hour at T=210°C. The resulting CF morphology of Fig.68 appeared to be completely dissolved, with the traps very distant among each other and a total of 182 O^-/V_O recombination events taking place either at the top interface and in the middle of the filament. As a consequence, the LRS cell state was not ensured anymore, since a resistance value of the order of $10^5\Omega$ was read in output.

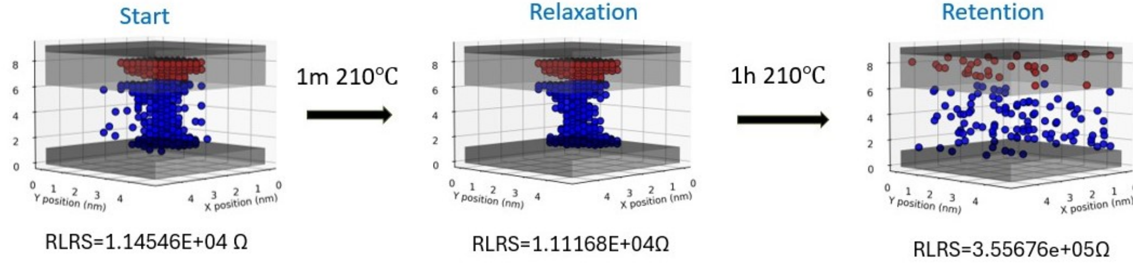


Figure 68: OxRAM cell #0 in his starting configuration (left), after 1m relaxation (middle) and after 1h 210°C bake (right)

It is evident, that such configuration is not stable, thus new attempts were undertaken.
In particular, to limit CF dissolution, the vacancy diffusion barrier was increased up to 2.3eV and, among the possible single-cell outputs reported in Fig.69, the $E_{A,vac} = 1.7eV$ option was selected, because to it corresponds the lowest R-drift.

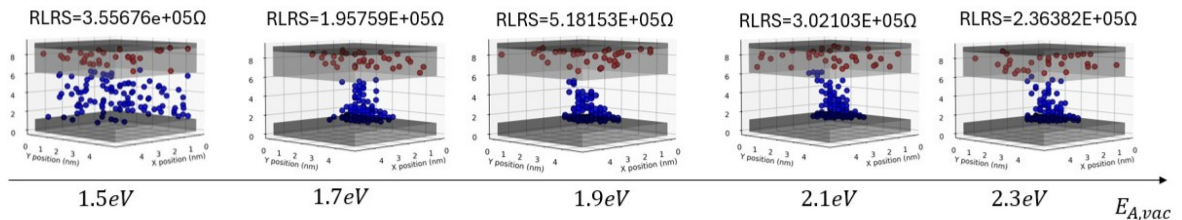


Figure 69: Effect of $E_{A,vac}$ on OxRAM cell #0 for 1h 210°C baking

Still the LRS condition is not ensured, due to the fact that recombination was the dominant effect, and, even for the 1.7eV case, a non-negligible gap of $g = 1.04nm$ was originated at the top interface, making conductivity poor.
The reason behind the high recombination probability could be addressed to a too low ion emission barrier from Ti to HfO_2. If its values is increased, ion emission is less prone to occur, together with recombination.
The results, in terms of cell morphology, recombination probabilities curves and ion and vacancy densities are reported in Fig.70.

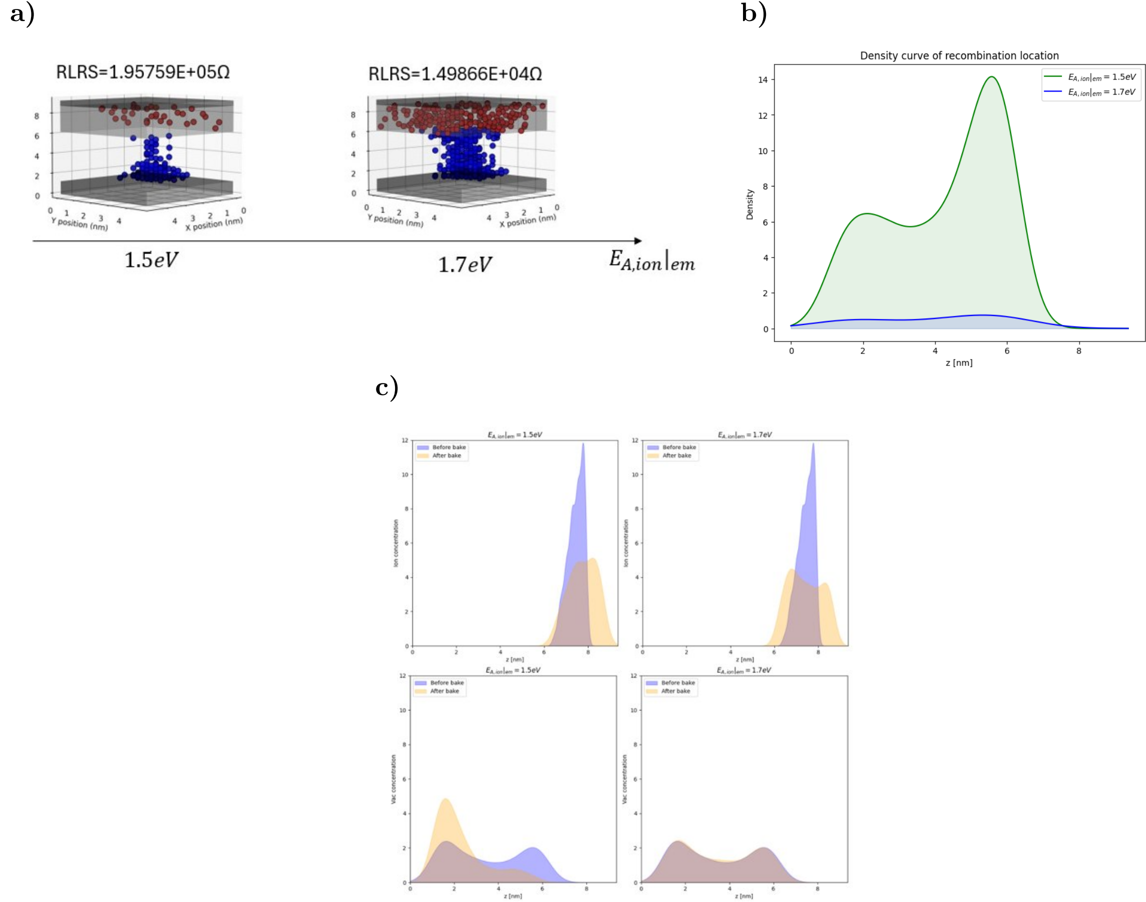


Figure 70: a) Snapshot cell #0 after 1h 210°C baking for $E_{A,ion}|_{em} = 1.5eV$ and $= 1.7eV$ b) Recombination density curve c) Ion and vacancy density along z-axis before and after bake.

It can be seen that for $E_{A,ion}|_{em} = 1.7eV$, the LRS of the memory was preserved, information data were retained, and recombination events were minimal, with only 15 events compared to the 181 in the 1.5eV scenario. This resulted in a thicker filament where more percolation paths were identified, thereby increasing overall conductivity.
This is indicative of the fact that, again, the interface region must be treated in a different way with respect to the whole scavenging layer.
At this point, the eligible energy configuration was found and it is resumed in Table16.

$E_{A,rec}$	$E_{A,ion}\|_{ox}$	$E_{A,ion}\|_{TiOx}$	$E_{A,ion}\|_{abs}$	$E_{A,ion}\|_{em}$	$E_{A,vac}$
0.2eV	0.2eV	1.5eV	1.1eV	1.7eV	1.7eV

Table 16: Activation energy setup of the eligible configuration.

A summary of the process flow is elucidated in the following Figures71, assuming as starting condition the "relaxed" cell of Fig.68.

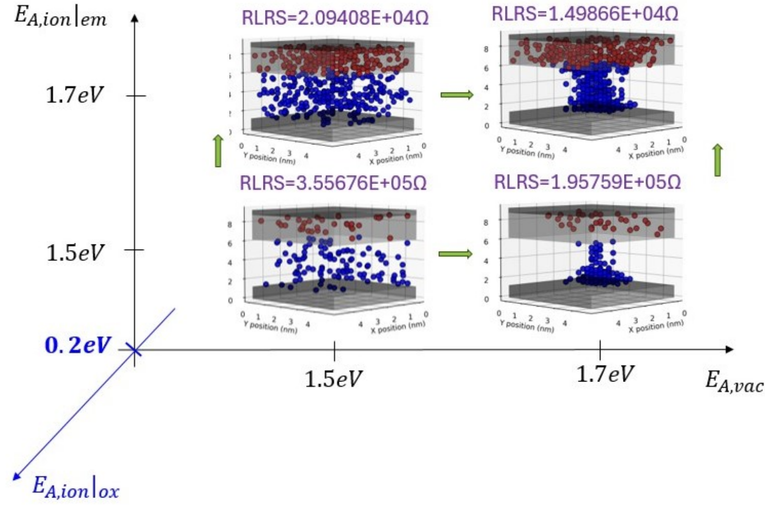


Figure 71: Effect of $E_{A,vac}$ and $E_{A,ion}|_{em}$ on the final snapshot of the cell.

Hence, for low (= 1.5eV) vacancy diffusion energy, dissolution of the filament is observed, while for high (= 1.7eV) energy, the CF is more compact. At the same time, for low (= 1.5eV) emission barrier, a lot of O^-/V_O recombinations take place, with further depletion of the TiO_x layer. Whereas for higher (= 1.7eV) ion emission barrier, recombination probability is deprimented.
The latter statement is confirmed by the recombination curves of Fig.72.

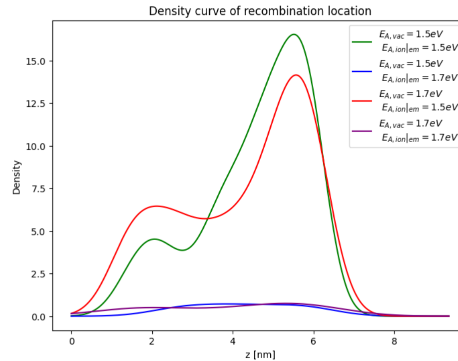


Figure 72: Recombination density curves for the 4 combinations of $E_{A,vac}$ and $E_{A,ion}|_{em}$, after 1h 210°C retention of cell #0.

Not only recombination probability, but also the trap-to-trap distance is important to assess which is the stable configuration. For this reason, considering a mesh cube having maximum $V_O - V_O$ distance

$= dx\sqrt{3} = 0.26\sqrt{3}nm$, the 1NN[4] occurance frequency was calculated either for low and high emission barrier.

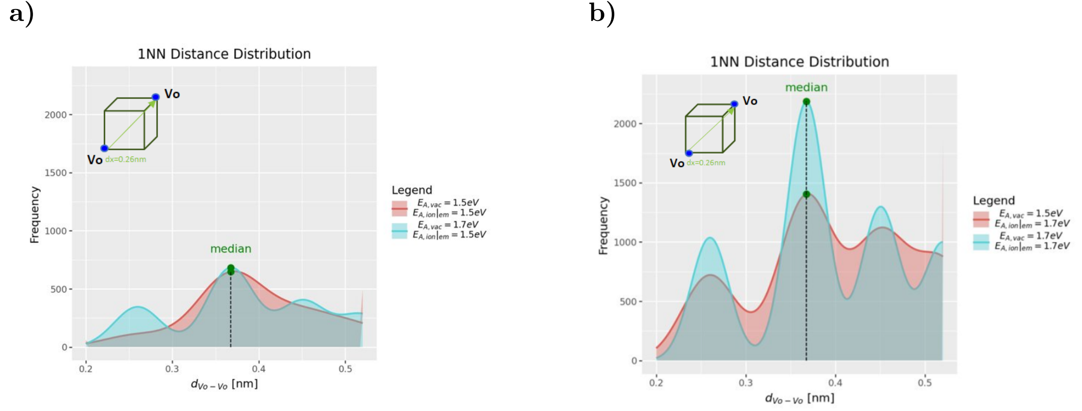


Figure 73: a) 1NN $V_O - V_O$ distance density for low (= 1.5eV) ion emission barrier. b) 1NN $V_O - V_O$ distance density for high (= 1.7eV) ion emission barrier.

Fig.73b) plays a crucial role in illustrating the reduction in resistance when moving from $E_{vac} = 1.5eV$ to $1.7eV$. Despite a higher trap density being observed within the resistive medium at lower V_O diffusion energy, the sharper density curve of the $1.7eV$ indicates closer traps. Based on the conductivity theory of Sec.3.2.5 and Fig.38, a metallic behavior exhibiting higher conductivity G is noticeable at shorter V_O distances a.
As a confirm of the enhanced stability for the 1.7eV-1.7eV case, the evolution of trap distance before and after retention is illustrated in Fig.74.

[4] nearest neighbour

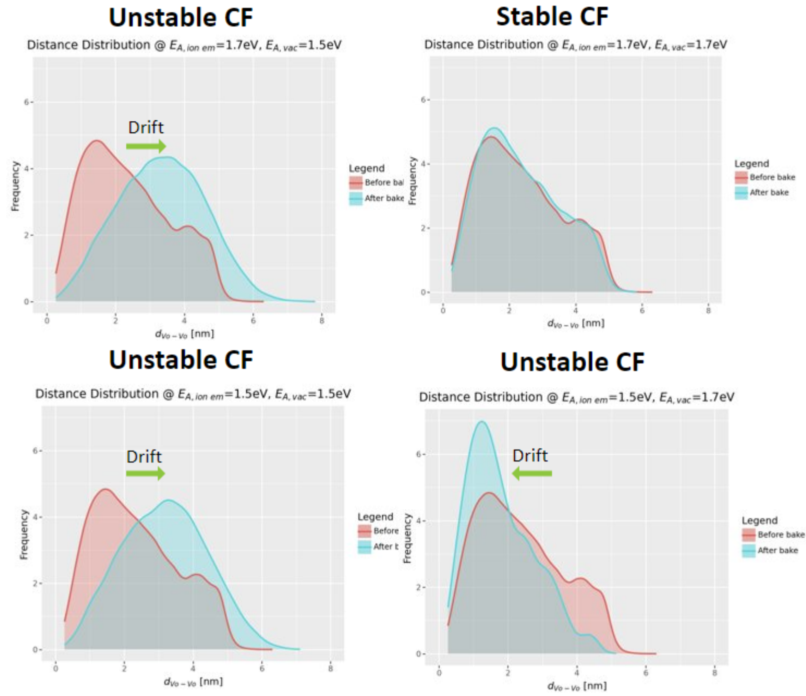


Figure 74: Trap distance evolution before and after 1h 210°C retention simulation, for the four $E_{A,vac} - E_{A,ion}|_{em}$ combinations.

Once that the energies were set according to Table16, retention simulations were carried out at four distinct temperatures, from 180°C up to 210°C, and multiple baking times, up to 5 hours. Examples of the 210°C (worst scenario) distribution and QQ-plot, for varying t_{bake}, are reported in the following.

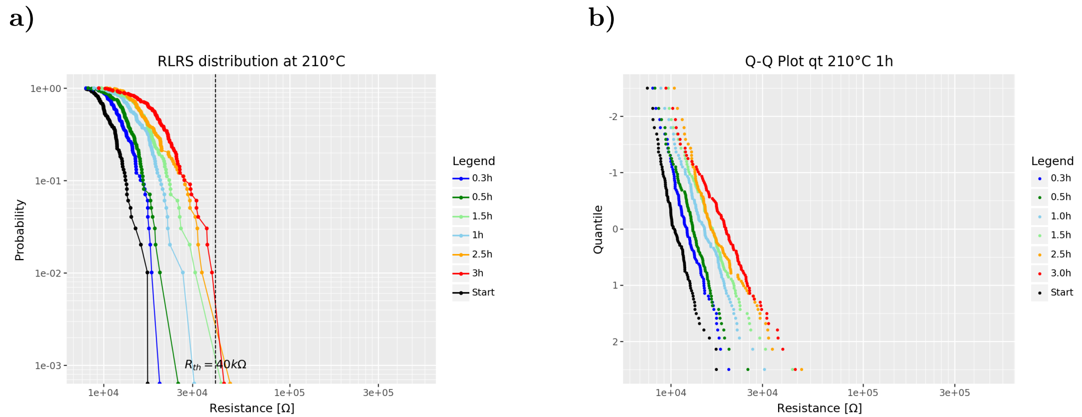


Figure 75: a) RLRS distribution and b) QQ-plot at 210°C for varying baking time. The maximum resistance registered is 48kΩ.

By choosing the same failure threshold of $R_{th} = 40k\Omega$, the fail probabilities $F(t, T)$ were extracted, following the same scheme of Par.4.4.6.

4.6.3 Weibull and TTF

According to the linearization equation 4.2 and to the Arrhenius formula 4.3, the 10 years lifetime was evaluated by looking at Weibull and TTF plots.

a) b)

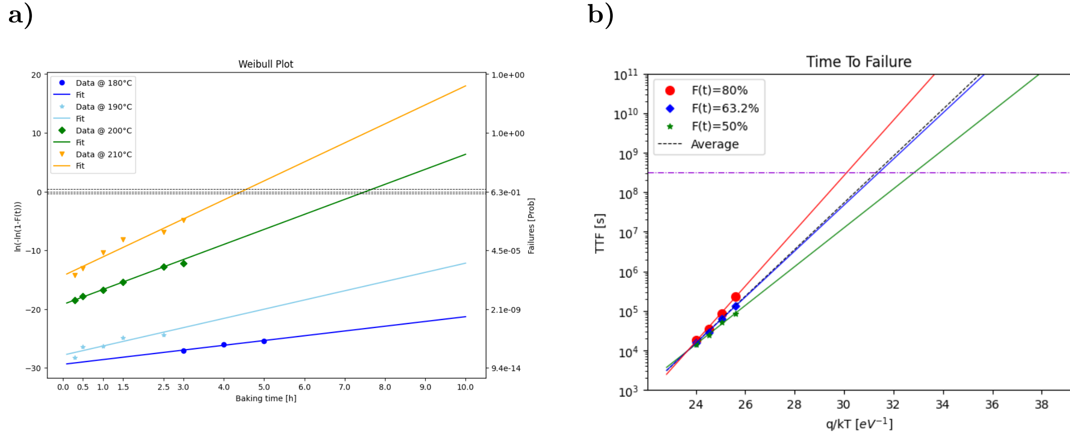


Figure 76: a) Weibull and b) TTF plot for the eligible configuration. The final slope is Ea=1.36eV with a 10ys maximum temperature of 98°C

Temperature[°C]	α[h]	β
180°C	36.2 h	0.81
190°C	17.8 h	1.57
200°C	7.52 h	2.56
210°C	4.4 h	3.23

Table 17: Scale (α) and shape (β) parameter for the 4 different Weibull lines

	$TTF_{180°C}$	$TTF_{190°C}$	$TTF_{200°C}$	$TTF_{210°C}$	E_A [eV]	T [°C]
50%	8.3×10^4s	5.1×10^4s	2.4×10^4s	1.4×10^4s	1.1eV	80°C
63.2%	1.3×10^5s	6.4×10^4s	2.7×10^4s	1.6×10^4s	1.34eV	96°C
80%	2.4×10^5s	8.7×10^4s	3.3×10^4s	1.8×10^4s	1.6eV	112°C

Table 18: Lifetimes, activation energy and maximum 10ys temperature extracted for three failing probabilities: 50%, 63.2%, 80%

Based on these results, several points must be clarified.

- The scale parameters α are larger with respect to the ion/interface case, and that is indicative of a more reliable device. For example, at 180°C, in order to reach the 63.2% probability of failure, it is needed a huge baking time of 36 hours.
- Also the shape parameters β are lower, but not the same. In fact, the Weibull lines are not parallel anymore and this could be a drawback since the physical mechanisms occurring within the cell during retention tests, are not predictable and change from one temperature to the other.

- Some Weibull data of the 210°C curve are misaligned with respect to the line. Indeed, a sort of ”saturation” behavior can be observed for higher times and temperatures, meaning that other phenomena take place in such conditions.
- The average slope of the TTF curves is $E_A = 1.36eV$ with a maximum operating temperature of 98°C.

The responsible of a not too high TTF slope is probably due to the saturation effect in the Weibull lines. Hence, to enforce such assumption, calculations were redone neglecting the data points 2.5h and 3h at 210°C.
The resulting plots are:

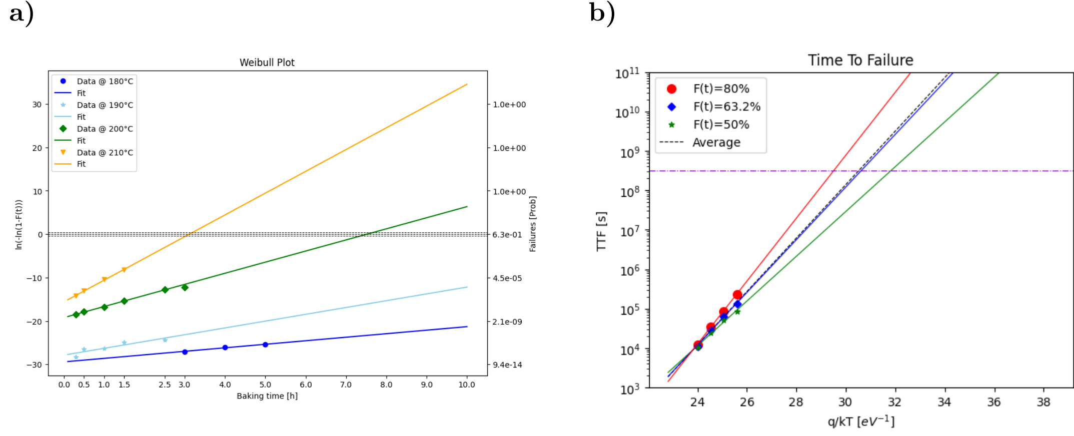


Figure 77: a) Weibull and b) TTF plot for the eligible configuration, neglecting the point 2.5h and 3h of the 210°C line. The final slope is now Ea=1.56eV with a 10ys maximum temperature of 106°C

Temperature[°C]	α[h]	β
180°C	36.2 h	0.81
190°C	17.8 h	1.57
200°C	7.52 h	2.56
210°C	3.1 h	5.01

Table 19: Scale (α) and shape (β) parameter for the 4 different Weibull lines

	$TTF_{180°C}$	$TTF_{190°C}$	$TTF_{200°C}$	$TTF_{210°C}$	E_A [eV]	T [°C]
50%	8.3×10^4s	5.1×10^4s	2.4×10^4s	1.1×10^4s	1.3eV	91°C
63.2%	1.3×10^5s	6.4×10^4s	2.7×10^4s	1.2×10^4s	1.53eV	106°C
80%	2.4×10^5s	8.7×10^4s	3.3×10^4s	1.3×10^4s	1.84eV	119°C

Table 20: Lifetimes, activation energy and maximum 10ys temperature extracted for three failing probabilities: 50%, 63.2%, 80%

The TTF slope is now $E_A = 1.56eV$, thus very close to the experimental value found in Fig.42b). Hereby, from a physical standpoint, the activation energy for the degradation process could be related to the ion diffusion energy inside TiO_x (see Table16), but it is also a combined effect of the interplay of the interface role and vacancy diffusion in the oxide.

Whereas, in terms of operating temperature, a maximum $T = 106°C$ was ensured for 10 years operation. Such value is enough to fulfill the AEC Grade-2 requirement, according to Table4.
Although the 150°C goal was not achieved, this configuration underscores a highly reliable NVM device, making it an useful basis for further tests and material engineering assumptions.

4.7 Single-cell LRS-retention with 3 V_O charge states

While approaching to the end of the Simulations process flow (Fig.47), a last bunch of tests were undertaken. Indeed, in order to mimic reality, the three charge states of an oxygen vacancy should be treated differently during ALTs.
The way V_O, V_O^+, and V_O^{2+} were discriminated according to the coordination number was already treated in Sec.4.5.
Successively, the same *ab initio* energy parameters of Fig.71, are employed for coherent comparison in Fig.78
Another difference to be outlined is that, with respect to the previous case, now, the starting condition was directly the cell after Forming programming, no relaxation phase was considered.

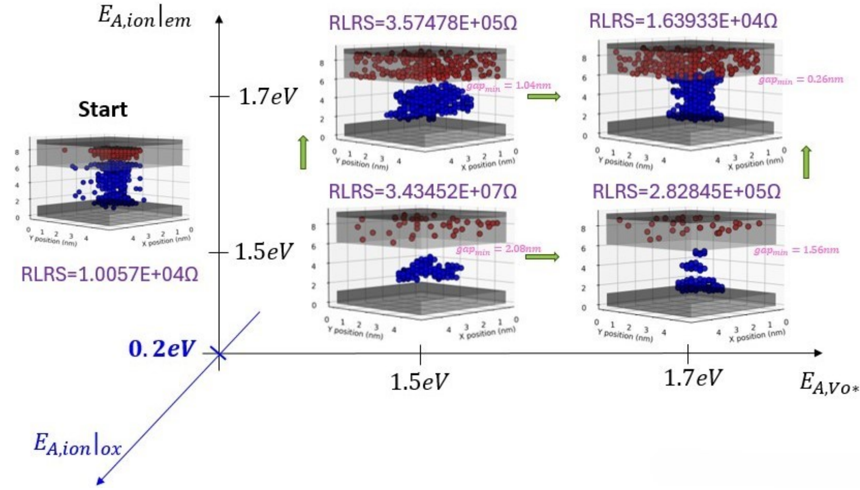


Figure 78: Effect of $E_{A,vac}$ and $E_{A,ion}|_{em}$ on the 1h 210°C retention of cell #0.$E_{A,vac}$ is the activation energy of neutral vacancy diffusion.

The other energy parameters are listed in Table21.

$E_{A,rec}$	$E_{A,ion}\|_{ox}$	$E_{A,ion}\|_{TiOx}$	$E_{A,ion}\|_{abs}$
0.2eV	0.2eV	1.5eV	1.1eV

Table 21: Activation energy setup.

In terms of resistance value, the highest conductivity was ensured for $E_{A,ion}|_{em} = 1.7eV$ and $E_{A,vac} = 1.7eV$, where, the latter refers to neutral vacancies only. Therefore:

$$V_O^+ \longrightarrow E_A = 0.8eV$$
$$V_O^{2+} \longrightarrow E_A = 0.7eV$$

Despite the simulation conditions and the energy parameters are the same used in Sec.4.6, at a first glance, the cell snapshots of Fig.78 appear very different. Indeed, as it was already seen in the first

relaxation phase, a "coalescence" of the CF is observed. What happened was that all the isolated V_O^{2+} recombined while V_O^{+} and V_O^{0} tended to associate and nucleate within the conductive region of Hafnia.
To be specific, when the vacancy diffusion was high (= 1.7eV), coalescence of the CF occured. In contrast, with lower diffusion values (= 1.5eV), there was a concentration of traps in the middle z region of the resistive medium, not consistent with what has been shown up to now. Similarly, a higher ion emission barrier (= 1.7eV) resulted in less frequent recombination events compared to a lower barrier (= 1.5eV), which saw over 200 recombination events throughout the entire CF length (Fig.79).

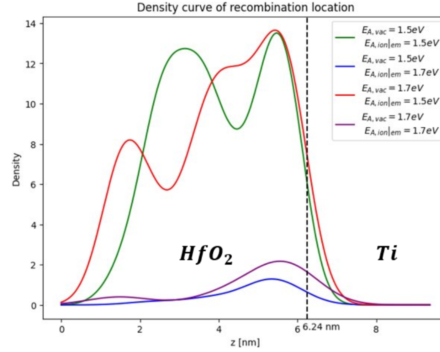


Figure 79: Recombination density curves for the 4 combinations of E_{A,V_O^0} and $E_{A,ion}|_{em}$ after 1h 210°C retention.

Notice that, due to a non-zero probability of recombination in the optimal configuration, the trap density within the active medium was reduced. This reduction explains the increase in resistance from the initial condition to the post-bake state, although the post-bake cell, at a first glance, appears to be more stable than the original one.

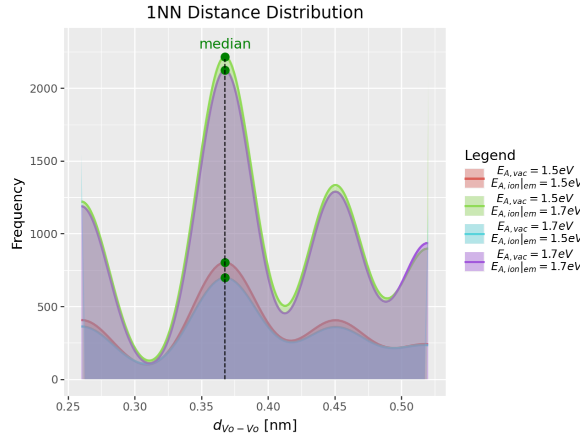


Figure 80: 1NN $V_O - V_O$ distance density for the 4 combinations of E_{A,V_O^0} and $E_{A,ion}|_{em}$ after 1h 210°C retention.

This time, by looking at the 1NN calculations of Fig.80 and according to what was affirmed previously, the most stable condition is expected to be the one having respectively $E_{A,vac} = 1.5eV$ and $E_{A,ion}|_{em} = 1.7eV$, since to that, is assigned the most peaked density curve. However, going back to the resistance values and to the snapshots of Fig.78, the LRS is not maintained, because of two

non-negligible gaps either at the top and bottom interfaces.
Therefore, merely examining the output filament morphology or the 1NN distance distribution is insufficient to determine if and why a particular configuration is stable. It is fundamental to combine multiple assumptions, calculations, and plots to make such a determination.
Finally, all the trap-to-trap distance combination are calculated and reported in Fig.81.

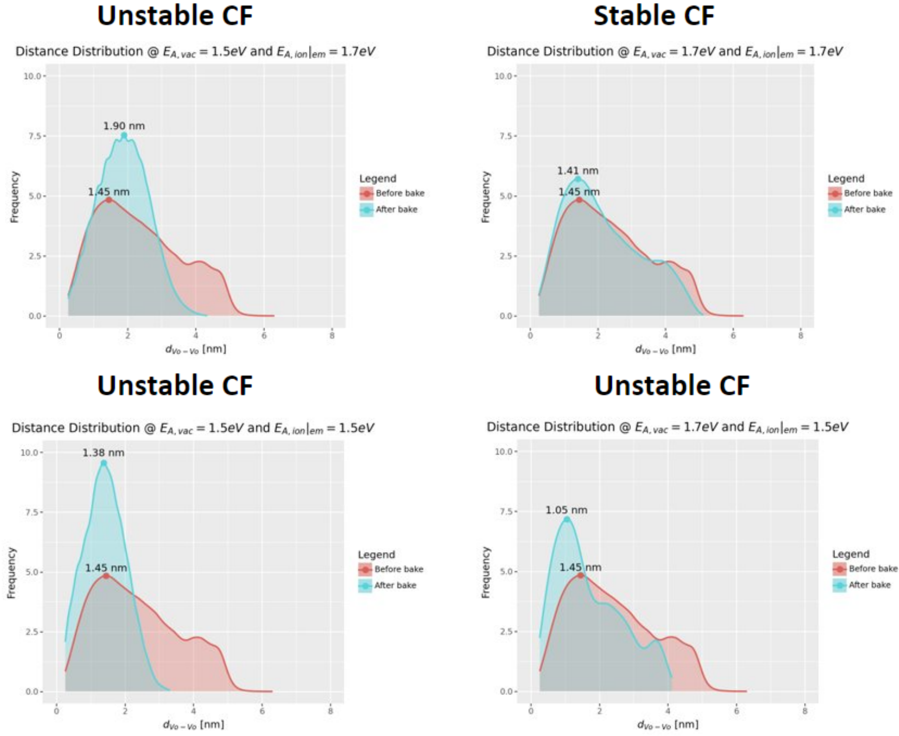


Figure 81: Trap distance evolution before and after 1h 210°C retention simulation for the 4 $E_{A,V_O^0} - E_{A,ion}|_{em}$ combinations.

For this third series of simulations, no further measures were carried out, as it is necessary to further investigate the cause of trap coalescence and the discrimination method of the V_O species before proceeding with multiple-cell calculations.

Reset Programming and HRS retention

The mechanisms involved during Reset programming are investigated in this last Chapter. In particular, it will be demonstrated that the HRS could be reached, regardless of whether O^{-}/V_O pair recombination effects take place or not.
Ultimately, single-cell HRS retention is evaluated for increasing baking time and fixed temperature.

Alongside with LRS, the stability of the programmed state has to be ensured over the years also for the High Resistance State (HRS). Indeed, according to the BER Theory illustrated in 4.1.2, information data is retained up to a certain LRS-HRS window margin.
To this aim, the 100 OxRAM cells exploited up to now were first programmed to the HRS and, subsequently, ALT simulations results are discussed.

5.1 Reset Programming

In order to re-program the OxRAM device into the Reset mode, what is done experimentally is to apply a reverse bias at the anode. In particular, according to Ref.[**9**], a larger reset voltage $|V_{stop}|$ result in a larger gap of the CF, hence, into a higher resistance R_{HRS}.
Furthermore, in terms of $I - V$ characteristic, since a TiN/Ti electrode couple is employed, a progressive reset transition is expected[34]. That is due to the Oxygen ion emission from the scavenging layer and gradual recombination with vacancies, as depicted in the insight of Fig.82.

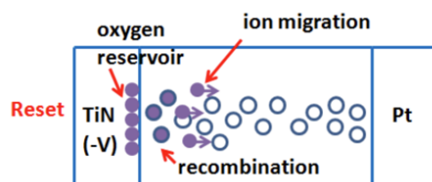


Figure 82: Schematics of Reset switching. Adapted from [**9**].

5.1.1 Parameters Setup

First the waveform of Reset pulse voltage is assessed. The amplitude voltage used during Reset programming simulation was $|V_{reset}| = 2.5V$, while the corresponding width saw: $t_{ramp\ up} \equiv t_{ramp\ dw} = 100ns$ and $t_{plateau} = 1\mu s$.

Thereafter, following the scheme of Table 3, many parameters were modified and listed in the following (Table 22).

Parameter	Value
V_{reset}	-2.5V
I_{comp}	1.2mA
$E_{A,R}$	0.1eV
$E_{A,ion}\|_{ox}$	0.2eV
$E_{A,ion}\|_{TiOx}$	variable
ε_{r,HfO_2}	16
ε_{r,TiO_x}	80
$\varepsilon_{r,metal}$	1024
$nmos_comp$	True

Table 22: Simulation initialization main parameters used for Reset programming.

Concerning vacancies, instead, a new approach was proposed to carry out meaningful results. The insulator medium was assumed to be divided into three regions of unequal z-length, as depicted in Fig.83

- A top and bottom region, each extending $3dx$ units, where only neutral V_O^0 were present. Here, the corresponding activation energy was set to $E_{A,V_O^0} = 1.4eV$.
- A middle, larger ($= 13dx$), region where doubly charged V_O^{2+} were present. The reason behind this has to be found in the studies of Ref.[**4**], which indicates that the central portion of the CF undergoes the highest temperature, making it more susceptible to CF breaking. The relative migration energy was assigned to $E_{A,V_O^{2+}} = 0.4eV$.

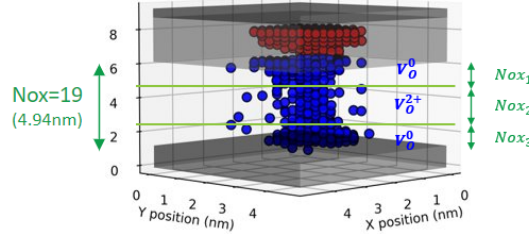


Figure 83: Initial LRS-programmed OxRAM cell where the subdivision of the resistive medium into three regions is outlined.

5.1.2 Reset: with and without recombination

One final aspect to be discussed is the diffusion energy of ions within the oxygen reservoir layer. This was set to "variable" in Table 22, since, in the last LRS retention results, $E_{A,ion}|_{TiOx}$ was predicted to be the dominant mechanism for resistance degradation. Moreover, in Ref.[**64**], it was discovered that the generation of a highly-resistive gap in the filament could be a result of either O^-/V_O pair recombination or the CF dissolution due to V_O spreading out of it.
Favouring one instead of the other would lead to different outcomes in the current evolution.

Following the KMC workflow (Fig.30) adopted for Forming operation, also for the Reset one, two distinct results were achieved depending on the $E_{A,ion}|_{TiOx}$ value.

- It was found that, for $E_{A,ion}|_{TiOx} = 0.6eV$, the Recombination case is verified. Indeed, for increasing simulation time, more and more recombination events were registered, due to a more favorable ion mobility and easiness of ejection from the TiO_x layer to the HfO_2 one. However, as it can be seen from the cell snapshot of Fig.84, it may happen that some ions cross the whole insulator thickness and reach the bottom electrode (TiN) interface without being recombined. The presence of these ions plays a crucial role during retention simulation, as it will be seen in Sec.5.2.

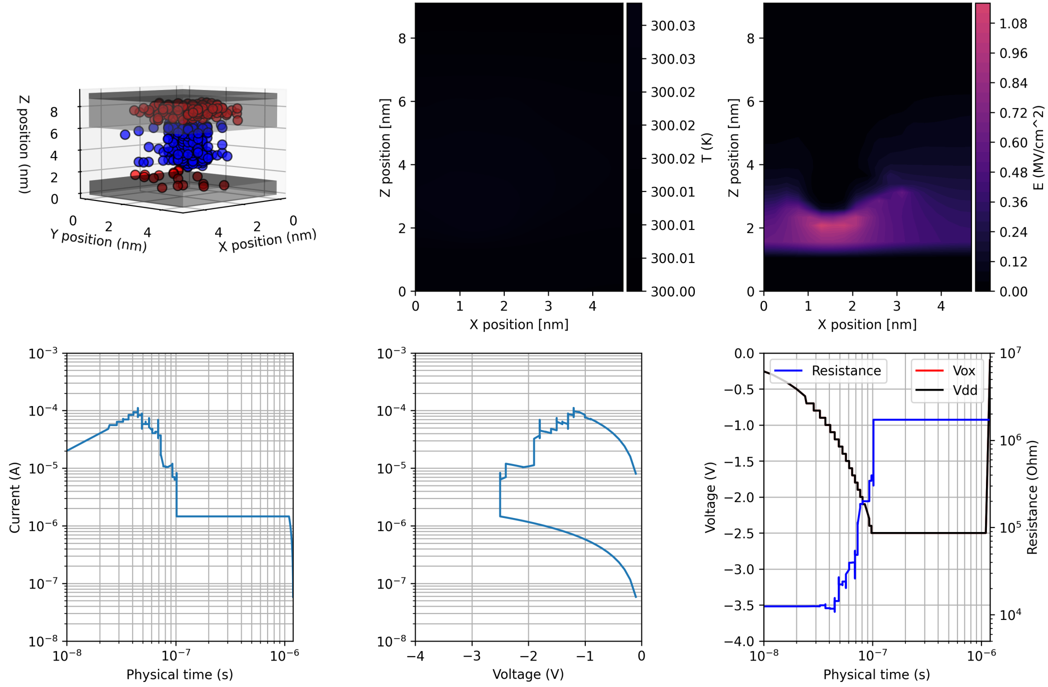


Figure 84: OxRAM cell after Reset Programming, Recombination case. In the top graphs are represented the local Temperature and Electric Field 2D maps, outlining the highest field concentration in correspondence of the tip of the CF. In the bottom graphs, from left to right are reported the evolution of current with time, the $I-V$ characteristic, and voltage (black) and resistance (blue) time behavior.

The output resistance value, after 0.1V reading was $R_{HRS} = 1.29882 \times 10^6\Omega$, ensuring an overall HRS/LRS ratio higher than 2 orders of magnitude.
In this context, it is fundamental to emphasize the evolution of current with time, as it exhibited an abrupt decrease when the V_{reset} pulse reached the plateau phase. This happened because the majority of recombination events took place during the initial ramp-up phase, and a $10^6\Omega$ resistance was already been achieved.
Such resistance jump was observable in a similar way for most of the programmed cells, whose single-resistance voltage behavior is depicted in Fig.85.

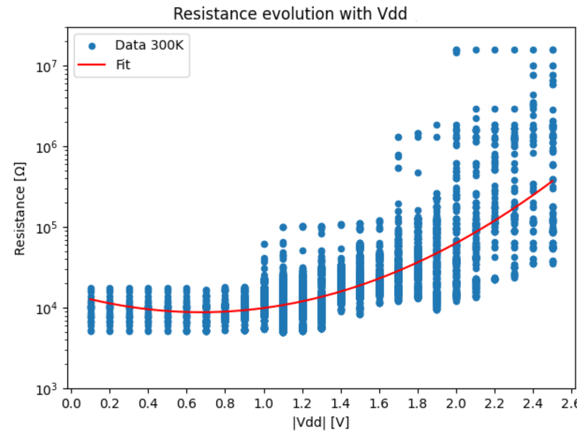


Figure 85: Single resistance evolution of multiple OxRAM cells during Reset programming in the ramp-up phase of the V_{reset} pulse.

Reset Programming was carried out at room temperature (300K), however an interesting study could be to see what happens for cryogenic temperatures.

- Conversely, when increasing the migration barrier to $E_{A,ion}|_{TiOx} = 0.8eV$, ions were less prone to diffuse and that resulted in a complete absence of recombination event. Nevertheless, the HRS could still be reached, due to filament dissolution as a consequence of Joule heating, confirming the studies of Ref.[**64**]. In fact, the temperature gradient caused the vacancies to diffuse out of the filament and, therefore, the current fluctuates as can be seen from the $I(t)$ curve of Fig.86.

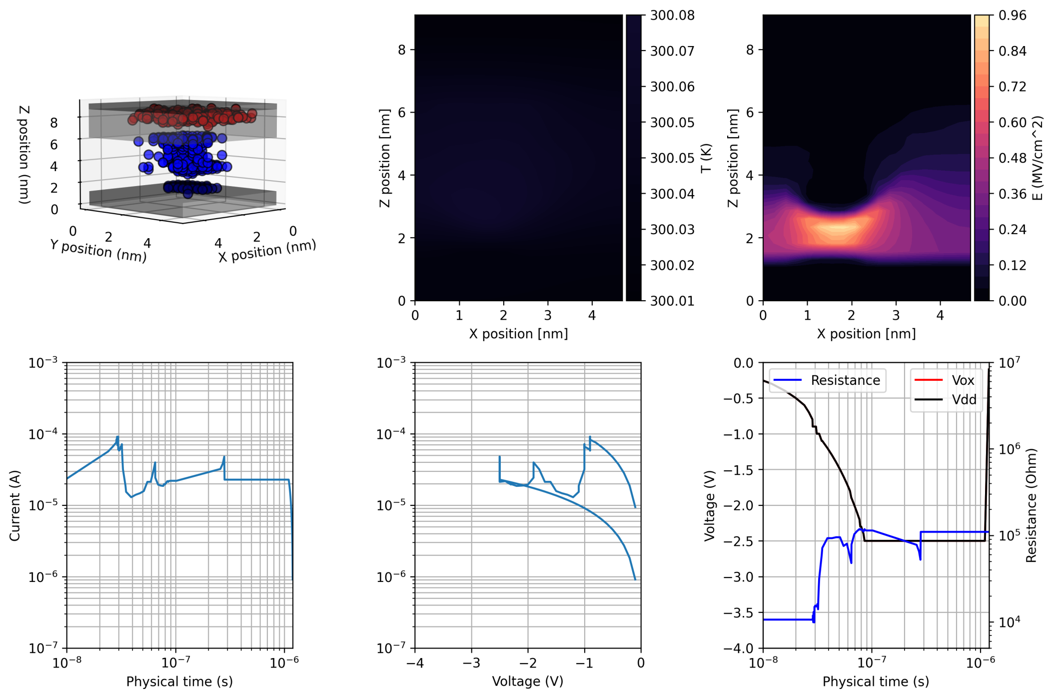


Figure 86: OxRAM cell in HRS without Recombination. In the top graphs, the local Temperature and Electric Field 2D maps. In the bottom graphs, from left to right, the evolution of current with time, the $I-V$ characteristic, and voltage (black) and resistance (blue) time behavior.

The final resistance is now: $R_{HRS} = 1.09815 \times 10^5 \Omega$, thus maintaining $R_{HRS}/R_{LRS} > 10$.

5.2 HRS Retention

Among the two options, the Recombination one was preferred, and retention simulations were carried out with the same eligible energy setup of Table16.
However, addressing an unique retention behavior of the HRS is not trivial, and it necessitates different evaluation methods compared to those of LRS[3]. For this reason, in this work, as the same tool employed so far in LRS was used, only single-cell retention simulations were conducted.
The aim of this procedure was to draw insights about the impact of long baking times and high temperature (fixed to 210°C).
In Figures 87-88 are reported the resistance evolution with time and the recombination density curves.

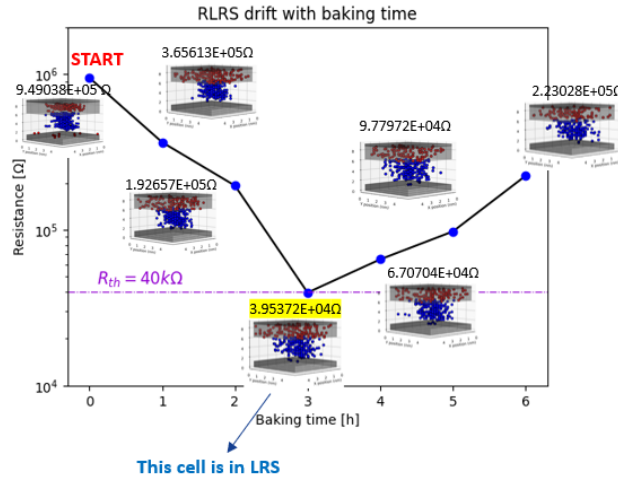


Figure 87: R_{HRS} and cell evolution with baking time, ranging from 1h to 6h, at T=210°C.

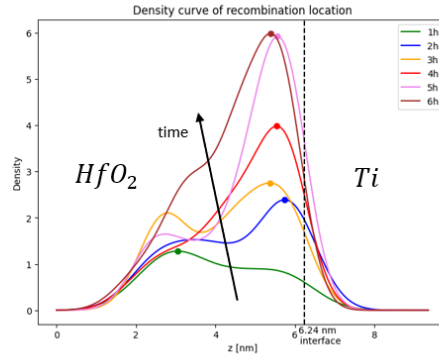


Figure 88: Recombination density curves for increasing baking time.

Pursuant to the outcomes of Fig.87, the OxRAM cell returned to the LRS for $t_{bake} = 3h$, because a resistance below 40kΩ was read. Nonetheless, this would not be true for all the 100 devices, because the stochasticity of events and of single KMC simulations must be always taken into account.
Additionally, for what regard recombination, the longer the bake, the more probable was its occurrence. It took place, mainly, at the top Ti/HfO_2 interface, except for the 1h retention, during which

all the oxygen ions deposited at the bottom electrode annihilated with the nearby vacancies.
For increasing baking time, another trend was observed when plotting the 1NN trap distance distribution (Fig.89). Here, the longer the duration, the less packed result to be the filament, increasing the number of voids between the possible percolation paths.

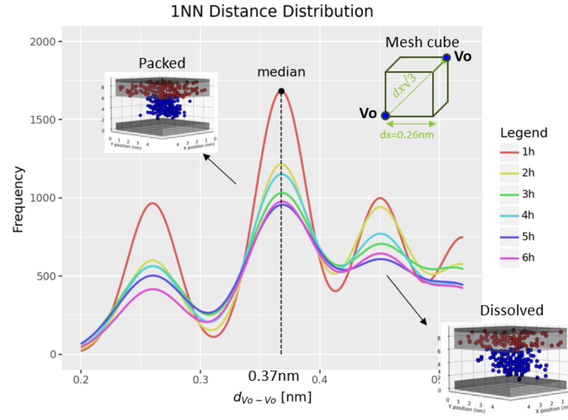


Figure 89: 1NN trap distance distribution of the HRS cells for increasing baking time during HRS retention at 210°C.

From merely examining the last two plots, one might assume that the lowest baking time guarantees the highest conductivity. However, as previously seen, the less resistive configuration was exhibited at 3 hours baking. The reason behind this is a still significant trap density within the oxide, together with a narrowing gap at the bottom electrode, as confirmed by the findings of Fig.90.

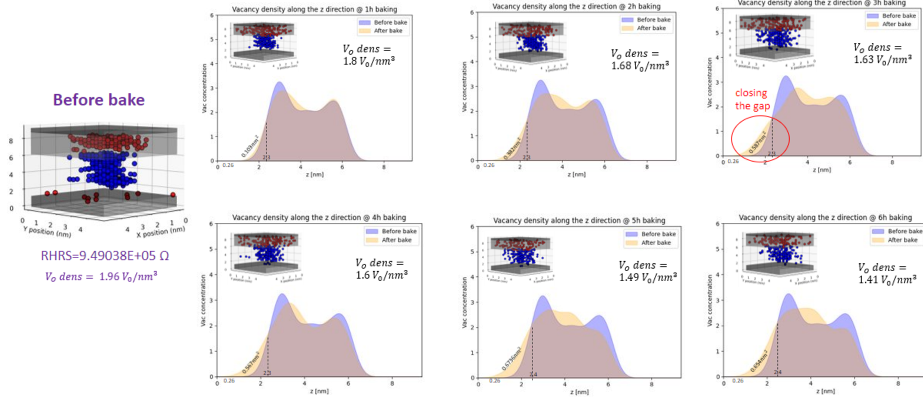


Figure 90: Trap density along the z-axis before and after bake for increasing time. The total V_O density is reported for the 6 cells, in conjunction with the V_O density close to the bottom electrode.

In conclusion, HRS retention needs much more instruments to be assessed, since either an increase or decrease of resistance, accompanied by fluctuations of the order of $\pm 10^3 \Omega$, could be expected according to the experimental distributions reported in Chap3.
The outcomes derived in the final segment of the present study could serve as a valuable starting point for further investigation of the simulation method implemented to replicate the ALT technique and the appropriate discretization of the MIM stack.

Conclusions and future perspectives

Approaching to the end of this work, it seeks to address the reliability of a ReRAM device in order to predict both the cell-to-cell and cycle-to-cycle variability when integrated into a larger crossbar memory array.
The typical approach exploited during the experimental characterization session is the well-known *Accelerated Life Test*. This technique is designed to measure the long-term reliability of the memory cell under elevated temperature conditions. By doing so, the failure mechanisms are accelerated, allowing to infer the device's performance over its operational lifespan.

To enforce the experimental data indicating an overall TTF slope of $E_A = 1.5eV$, a Python-based model was developed using the same methodology. Here, a Kinetic Monte Carlo framework was implemented to simulate the stochasticity of phenomena within a single OxRAM cell. Multiple simulations were conducted on a sample of 100 LRS-programmed RRAMs at varying temperatures and times.
In particular, to provide insights into the material engineering of the MIM stack, the activation energies for the diffusion of oxygen ions and vacancies within the cell were thoroughly determined. It was discovered that both the non-stoichiometric interface between the oxygen reservoir layer and the resistive medium, and the diffusion of V_O into the latter, are crucial for conducting a comprehensive analysis of filament dissolution and its capability of retaining data.
In this scenario, starting from *ab initio*-related parameters, the optimal energy configuration that minimized failure probabilities was identified. Specifically, a good window margin was ensured up to 210°C-3 hours baking, and distinct phenomena were observed in the Weibull plot across various temperatures. The final TTF line slope was determined to be 1.56eV, consistent with experimental data, with the failure origin being a result of the interplay of Oxygen ion diffusion within TiO_x, Vacancy diffusion within HfO_2 and interface effects. These findings should be proven by further electrical characterization tests.

In addition, to present a more "realistic" scenario, this work assumes monoclinic Hafnia as the active medium throughout. However, the inclusion of some doping could be considered. Furthermore, while the oxygen vacancy diffusion was assumed isotropic, it is established that this diffusion varies with the crystal lattice orientation and is notably favored along grain boundaries, where the structural disorder allows for easier defect diffusion. To mimic CF dissolution, it is essential to further investigate how the three charge states and the ionization mechanism are addressed.

Likewise, HRS retention showed promising results and practical insights for future dissertations. However, other considerations must be done when modeling HRS retention. For instance, ions at the bottom electrode might either recombine with vacancies, so as to rupture the filament, or being trapped at the bottom electrode interface, thereby maintaining the gap region. This latter scenario is speculative, based on the observation that TiN has a non-negligible, albeit lower electronegativity than Ti.

Bibliography

[1] Blocks & Files "*Resistance is not futile, according to ReRAM startup Weebit Nano*" [Online] `https://blocksandfiles.com/2021/04/26/resistance-is-futile-not-so-says-reram-startup-weebit-nano/`

[2] IDC, und Statista. *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 (in zettabytes).* Chart. June 7, 2021. Statista. Accessed May 07, 2024.

[3] N. Kopperberg, D. J. Wouters, R. Waser, S. Menzel, S. Wiefels; *Accurate evaluation method for HRS retention of VCM ReRAM. APL Mater.* 1 April 2024; 12 (3): 031112.

[4] Daniele Ielmini 2016 *Resistive switching memories based on metal oxides: mechanisms, reliability and scaling* Semicond. Sci. Technol. 31 063002

[5] IRDS. *Beyond CMOS and Emerging Materials Integration.* IEEE, New York 2022.

[6] IRDS. *More than Moore.* IEEE, New York 2022.

[7] Pan, W., Li, Z., Zhang, Y. et al. *The New Hardware Development Trend and the Challenges in Data Management and Analysis.* Data Sci. Eng. 3, 263–276 (2018)

[8] Lim K, Chang J, Mudge T *et al. Disaggregated memory for expansion and sharing in blade servers* 2009, ACM SIGARCH Comput Archit News 37(3):267–278

[9] S. Yu, Ximeng Guan and H. . -S. P. Wong, *On the stochastic nature of resistive switching in metal oxide RRAM: Physical modeling, monte carlo simulation, and experimental characterization* 2011 International Electron Devices Meeting, Washington, DC, USA

[10] D. Duncan, B. Magyari-Köpe and Y. Nishi, *Filament-Induced Anisotropic Oxygen Vacancy Diffusion and Charge Trapping Effects in Hafnium Oxide RRAM,* in IEEE Electron Device Letters, vol. 37, no. 4, pp. 400-403, April 2016, doi: 10.1109/LED.2016.2524450.

[11] Circuit Globe. *Difference Between Von Neumann and Harvard Architecture.* [Online]. Available: `https://circuitglobe.com/difference-between-von-neumann-and-harvard-architecture.html`.

[12] Server LIFT. *9 Challenges That Data Centers Are Facing Now.* [Online]. Available: `https://serverlift.com/blog/data-center-challenges/`.

[13] Institute for Microelectronics at Tu Wien. *Flash Memory* [Online] `https://www.iue.tuwien.ac.at/phd/windbacher/node14.html`

[14] Wikipedia. *Multi-level cell* [Online] `https://en.wikipedia.org/wiki/Multi-level_cell`

[15] Hamzah A., Hilman A., Tan M., Alias N., Johari Z., Ismail R. *Scaling Challenges of Floating Gate Non-Volatile Memory and Graphene as the Future Flash Memory Device: A Review.* Journal of Nanoelectronics and Optoelectronics. (2019) 14. 1195-1214. 10.1166/jno.2019.2204.

[16] Yale University. *Flash Drives: methods and materials* [Online] https://volga.eng.yale.edu/teaching-resources/flash-drives/methods-and-materials

[17] Chen, An. *A review of emerging non-volatile memory (NVM) technologies and applications.* Solid-state Electronics 125 (2016): 25-38.

[18] Fujitsu Electronics Europe, Factsheet. *Non-Volatile Random-Access Memory FRAM (Ferroelectric RAM)* [Online] https://www.fujitsu.com/uk/Images/C29-FRAM-Factsheet-170802-0959.pdf

[17] Ramtron *FRAM Technology Backgrounder. An overview of FRAM Technology* – Updated Dec. 2000

[18] R. Sbiaa, H. Meng, S. N. Piramanayagam *Materials with perpendicular magnetic anisotropy for magnetic random access memory.* 4 October 2011, Physica Status Solidi. https://doi.org/10.1002/pssr.201105420

[19] Q. Dong et al., *A 1Mb 28nm STT-MRAM with 2.8ns read access time at 1.2V VDD using single-cap offset-cancelled sense amplifier and in-situ self-write-termination,* 2018 IEEE International Solid-State Circuits Conference - (ISSCC), San Francisco, CA, USA, 2018, pp. 480-482, doi: 10.1109/ISSCC.2018.8310393.

[20] Dmytro Apalkov, Bernard Dieny, J. M Slaughter. *Magnetoresistive Random Access Memory.* Proceedings of the IEEE, 2016, 104, pp.1796 - 1830. ff10.1109/JPROC.2016.2590142ff. ffhal-01834195f

[21] Gupta, V., Kapre, A., Dubey, S.K., Islam, A. (2023). *Implementation and Analysis of CNFET-Based PCRAM Cell Using 32 nm Technology.* ESDA 2021. Lecture Notes in Electrical Engineering, vol 1057. Springer, Singapore. https://doi.org/10.1007/978-981-99-3691-5_21

[22] Paul Boldt, Semiconductor Insights *Patents eye phase-change RAM* EE Times, 2007. [Online] https://www.eetimes.com/patents-eye-phase-change-ram/?_ga

[23] Paul Boldt *numonyx – was that PCM or PCRAM?.* getting technology right, 2008. [Online]https://www.gettingtechnologyright.com/numonyx-was-that-pcm-or-pcram/

[24] Weebit Nano "*Meet the Weebit ReRAM Bitcell*".[Online] https://www.weebit-nano.com/technology/reram-bitcell/

[25] National Library of Medicine *Brain Power.* National Center for Biotechnology information, 2021. [Online] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8364152/

[26] Sangwan, V.K., Hersam, M.C. *Neuromorphic nanoelectronic materials.* Nature Nanotechnology. 15, 517–528 (2020). https://doi.org/10.1038/s41565-020-0647-z

[27] S. Yeluri, *Tearing down the memory wall.* Juniper Networks, 2022. [Online] https://community.juniper.net/blogs/sharada-yeluri/2022/08/22/tearing-down-memory-wall

[28] Leon O.Chua, *Memristor-The Missing Circuit Element.* IEEE transactions on circuit theory, vol. ct-18, no. 5, September 1971

[29] V. Gupta, G. Lucarelli, S. Castro-Hermosa,T. Brown, M. Ottavi. *Perovskite based Low Power Synaptic Memristor Device for Neuromorphic application.* May 2019. 10.1109/DTIS.2019.8734983.

[30] R. Waser, M. Aono *Nanoionics-based resistive switching memories.* Nature Mater 6, 833–840 (2007). https://doi.org/10.1038/nmat2023

[31] A. Sawa *Resistive switching in transition metal oxides*, Materials Today, 2008.

[32] E. Briman *The Power of ReRAM for PMICs* [Online] Weebit Nano, 2022. https://www.weebit-nano.com/the-power-of-reram-for-pmic-rram-memory-embedded-nvm/

[33] E. Ambrosi, A. Bricalli, M. Laudato, D. Ielmini *Impact of oxide and electrode materials on the switching characteristics of oxide ReRAM devices* Faraday Discuss., 2019, 213, 87

[34] B. Traoré, P. Blaise, E. Vianello, L. Perniola, B. De Salvo and Y. Nishi, "*HfO2-Based RRAM: Electrode Effects, Ti/HfO2 Interface, Charge Injection, and Oxygen (O) Defects Diffusion Through Experiment and Ab Initio Calculations,*" in IEEE Transactions on Electron Devices, vol. 63, no. 1, pp. 360-368, Jan. 2016, doi: 10.1109/TED.2015.2503145.

[35] C. Nail et al., "*Understanding RRAM endurance, retention and window margin trade-off using experimental results and simulations*," 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2016, pp. 4.5.1-4.5.4, doi: 10.1109/IEDM.2016.7838346.

[36] G. Molas, E. Nowak. "*Advances in Emerging Memory Technologies: From Data Storage to Artificial Intelligence.*" Applied Sciences, 2021. 11. 11254. 10.3390/app112311254.

[37] G. Molas, G. Piccolboni, A. Bricalli, A. Verdy, I. Naot, et al.. "*High temperature stability embedded ReRAM for 2x nm node and beyond.*" 2IEEE International Memory Workshop (IMW), May 2022, Dresden, France. pp.1-4, ⟨10.1109/IMW52921.2022.9779293⟩. ⟨hal-03840618⟩

[38] Weebit Nano. "*Where Weebit ReRAM Can Make a Difference*" [Online] https://www.weebit-nano.com/market/applications/

[39] H. Wang, "*Challenges in Automotive Memory Solutions*," 2018 IEEE International Memory Workshop (IMW), Kyoto, Japan, 2018, pp. 1-7, doi: 10.1109/IMW.2018.8388782.

[40] Weebit Nano "*Plotting the Course for ReRAM in Automotive*" EDTM 2024. [Online] https://www.weebit-nano.com/wp-content/uploads/2024/03/EDTM2024-Weebit-ReRAM-embedded-RRAM-IP-NVM-for-semiconductors-replace-Flash-memory.pdf

[41] P. Huang Peng, Y. Zhao, J. Kang . "*Resistive-Switching Memories.*" Springer Handbook of Semiconductor Devices (pp.1043-1092), 2022. 10.1007/978-3-030-79827-7_29.

[42] F. Zahoor, T. Zulkifli, F. Khanday, "*Resistive Random Access Memory (RRAM): an Overview of Materials, Switching Mechanism, Performance, Multilevel Cell (mlc) Storage, Modeling, and Applications.*" Nanoscale Res Lett 15, 90 (2020). https://doi.org/10.1186/s11671-020-03299-9

[43] M. Lanza, H.-S. P. Wong, E. Pop, D. Ielmini,*et al.* "*Recommended Methods to Study Resistive Switching Devices*"Adv. Electron. Mater. 2019, 5, 1800143. https://doi.org/10.1002/aelm.201800143

[44] Reganaz, L., Esmanhotto, E., Ait Abdelkader, N., Minguet Lopez, J., Castellani, N., Rafhay, Q., Deleruyelle, D., Grenouillet, L., Aussenac, F., Vianello, E., Andrieu, F. and Molas, G. (2022), "*Elucidating Postprogramming Relaxation in Multilevel Cell-Resistive Random Access Memory by Means of Experimental and Kinetic Monte Carlo Simulation* ". Data. Phys. Status Solidi A, 219: 2100753. https://doi.org/10.1002/pssa.202100753

[45] Ielmini, D., Milo, V. "*Physics-based modeling approaches of resistive switching devices for memory and in-memory computing applications.*" J Comput Electron 16, 1121–1143 (2017). https://doi.org/10.1007/s10825-017-1101-9

[46] Reganaz, L. *Integration of resistive memory in FDSOI 28nm node for In Memory Computing applications.* Université Grenoble Alpes, 2023.

[47] J. F. Kang et al., "*Modeling and design optimization of ReRAM,*" The 20th Asia and South Pacific Design Automation Conference, Chiba, Japan, 2015, pp. 576-581, doi:10.1109/ASPDAC.2015.7059070.

[48] N. Capron, P. Broqvist, A. Pasquarello, "*Migration of oxygen vacancy in HfO2 and across the HfO2/SiO2 interface: A first-principles investigation*"Appl. Phys. Lett. 2007, 91, 192905.

[49] L. Reganaz et al., "*Investigation of resistance fluctuations in ReRAM: physical origin, temporal dependence and impact on memory reliability*". 2023 IEEE International Reliability Physics Symposium (IRPS), Monterey, CA, USA, 2023, pp. 1-6,doi: 10.1109/IRPS48203.2023.10117882.

[50] J. Guy, G. Molas,P. Blaise, *et al.* "*Investigation of Forming, SET, and Data Retention of Conductive-Bridge Random-Access Memory for Stack Optimization*" (2015) IEEE Transactions on Electron Devices. 62. 1-1. 10.1109/TED.2015.2476825.

[51] G. Bersuker, D. Gilmer, D. Veksler,*et al.*"*Metal Oxide RRAM Switching Mechanism Based on Conductive Filament Properties.*" (2012). Journal of Applied Physics. 110. 124518. 10.1063/1.3671565.

[52] G. Milano *et al.* "*Quantum Conductance in Memristive Devices: Fundamentals, Developments, and Applications*" 2022, Advanced Materials. `https://doi.org/10.1002/adma.202201248`

[53] R. Landauer, "*Spatial Variation of Currents and Fields Due to Localized Scatterers in Metallic Conduction,*" IBM Journal of Research and Development, vol. 1, no. 3, pp. 223–231, Jul. 1957, conference Name: IBM Journal of Research and Development.

[54] E. Pérez, D. Maldonado, C. Acal, J. E. Ruiz-Castro, A. M. Aguilera, F. Jiménez-Molinos, J. B. Roldán, and C. Wenger, "*Advanced temperature dependent statistical analysis of forming voltage distributions for three different HfO2-based RRAM technologies,*" 2021, Solid-State Electronics, vol. 176, p. 107961.

[55] B. Govoreanu, S. Clima, I. P. Radu, Y.-Y. Chen, D. J. Wouters, and M. Jurczak, "*Complementary Role of Field and Temperature in Triggering ON/OFF Switching Mechanisms in Resistive RAM Cells,*" 2013. IEEE Transactions on Electron Devices, vol. 60, no. 8, pp. 2471–2478.

[56] L. Vandelli, A. Padovani, G. Bersuker, D. Gilmer, P. Pavan and L. Larcher, "*Modeling of the Forming Operation in HfO2-Based Resistive Switching Memories*," 2011 3rd IEEE International Memory Workshop (IMW), Monterey, CA, USA, 2011, pp. 1-4, doi: 10.1109/IMW.2011.5873224.

[57] V. G. Karpov, "*Adiabatic theory of SET and RESET transitions,*" 2021 Journal of Applied Physics, vol. 129, no. 11, p. 114501

[58] Z. Wei *et al.*, "*Demonstration of high-density ReRAM ensuring 10-year retention at 85°C based on a newly developed reliability model*". 2011 International Electron Devices Meeting, Washington, DC, USA, 2011, pp. 31.4.1-31.4.4, doi: 10.1109/IEDM.2011.6131650.

[59] E. Perez *et al.*, "*Data retention investigation in Al:HfO_2-based resistive random access memory arrays by using high-temperature accelerated tests*". 2019, Journal of Vacuum Science & Technology B 37, 012202; doi: 10.1116/1.5054983

[60] B. Cojocaru, D. Avram, R. Negrea, C. Ghica, VG. Kessler, GA. Seisenbaeva, VI. Parvulescu VI, C. Tiseanu, "*Phase Control in Hafnia: New Synthesis Approach and Convergence of Average and Local Structure Properties*". 2019, ACS Omega 22;4(5):8881-8891. doi: 10.1021/acsomega.9b00580. PMID: 31459976; PMCID: PMC6648616.

[61] K. Olmstead "*Challenges For Achieving Automotive Grade 1/0 Reliability In FCBGA and fcCSP Packages*". 2022, Semiconductor Engineering [Online] https://semiengineering.com/challenges-for-achieving-automotive-grade-1-0-reliability-in-fcbga-and-fccsp-packages/

[62] C. Wang, H. Wu, B. Gao, T. Zhang, Y. Yang, H. Qian, "*Conduction mechanisms, dynamics and stability in ReRAMs*". 2018 Microelectronic Engineering, Volumes 187–188, pp. 121-133, ISSN 0167-9317, https://doi.org/10.1016/j.mee.2017.11.003.

[63] S. Yu, C. Yin, *et.al* "*A Monte Carlo study of the low resistance state retention of HfO x based resistive switching memory*". 2012, Applied Physics Letters, 100(4), Article 043507. https://doi.org/10.1063/1.3679610

[64] X. Xu, B. Rajendran and M. P. Anantram, "*Kinetic Monte Carlo Simulation of Interface-Controlled Hafnia-Based Resistive Memory,*" in IEEE Transactions on Electron Devices, vol. 67, no. 1, pp. 118-124, Jan. 2020, doi: 10.1109/TED.2019.2953917.

[65] Wikipedia "*Weibull distribution*" [Online] https://en.wikipedia.org/wiki/Weibull_distribution

[66] Hang Meng, Shihao Huang, Yifeng Jiang. "*The role of oxygen vacancies on resistive switching properties of oxide materials*". AIMS Materials Science, 2020, 7(5): 665-683. doi: 10.3934/matersci.2020.5.665

[67] Intel. "*3D XPoint: A Breakthrough in Non-Volatile Memory Technology*" [Online] https://www.intel.com/content/www/us/en/architecture-and-technology/intel-micron-3d-xpoint-webcast.html

[68] T Francois, J Coignus, A Makosiej, B Giraud, C Carabasse, *et al* "*16kbit HfO2: Si-based 1T-1C FeRAM Arrays Demonstrating High Performance Operation and Solder Reflow Compatibility*". 67th Annual IEEE International Electron Devices Meeting (IEDM) 2021, Dec 2021, San Francisco, USA.

[69] G. Florio, "*Applications of Magnetic Materials*", Encyclopedia of Smart Materials, Elsevier, 2022, Pages 24-31, https://doi.org/10.1016/B978-0-12-815732-9.00067-X

[70] Luc Thomas, *et al* "*Perpendicular spin transfer torque magnetic random access memories with high spin torque efficiency and thermal stability for embedded applications*" J. Appl. Phys. 7 May 2014; 115 (17): 172615. https://doi.org/10.1063/1.4870917

[71] Mejri, I.H., Omri, K., Ghiloufi, I. et al. Resistive switching behavior in ZnO:Ca thin films deposited by a pulsed laser deposition technique. Appl. Phys. A 129, 210 (2023). https://doi.org/10.1007/s00339-023-06508-1