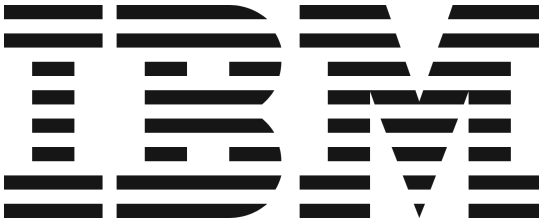


Politecnico di Torino

Department of Electronics and Telecommunications engineering,
MSc. in Nanotechnologies for Smart and Integrated Systems

Physics-Aware Compact Modeling of Analog Conductive-Metal-Oxide/HfO_x ReRAM Device

Master's Thesis - IBM Research Europe, Zurich

Matteo Galetta - s305915

Internal supervisor - Politecnico di Torino:
Prof. Carlo Ricciardi, Department of Applied Science and Technology

Local supervisors - IBM Research Europe:
Dr. Valeria Bragaglia, Neuromorphic Devices and Systems group
MSc. Donato Francesco Falcone, Neuromorphic Devices and Systems group

A.Y. 2023/2024

Abstract

Over the last decade, standard computing architectures based on von Neumann paradigm struggled to manage Internet of Things and Artificial Intelligence (AI) workloads. The inherently inefficient data transfer between processing and storage units is not suited to deal with modern data-centric applications, which are growing in complexity and scale. To tackle the exponentially increasing power demand of Neural Networks computing tasks, new low-power hardware implementations are necessary. In-Memory computing has the potential to fulfill the energy requirements in the modern Information Technology field, enabling parallel data processing and reducing latency. Memristive devices within cross-point architectures turned out to be a promising solution to perform analog In-Memory Computing, allowing to map multi-level weights between Neural Networks layers. Especially for training in neuromorphic hardware, Resistive Random Access Memory (ReRAM) devices attracted significant attention, offering fast and low-power switching capabilities, high scalability and non-volatile data storage.
The integration of ReRAMs in neuromorphic systems requires extensive optimization of several aspects of the technology, ranging from device materials and fabrication processes to physical/electrical features improvements. For these purposes, physical modeling becomes crucial to provide a detailed understanding and accurate predictivity of device performances.
At IBM Research Europe, the Neuromorphic Devices and Systems group is conducting R&D projects concerning the development and the optimization of an innovative ReRAM technology, integrated in system-level crossbar arrays for AI accelerators.
A robust compact model able to accurately capture the operation of ReRAM devices is essential to accelerate circuit-level simulations of memory arrays and neuromorphic hardware. Hereby this dissertation presents the development and the validation of a compact model for analog filamentary Conductive-Metal-Oxide/HfO_x ReRAM IBM technology. The model integrates a physics-based approach to describe analytically the ion migration mechanisms causing restistive switching phenomena. Further analysis concerns the switching dynamics of the device, evaluating the time scales in which resistive switching occurs. The model validation is conducted against experimental data of electrical characterizations, demonstrating the model's accuracy, robustness and the capability to capture the analog behavior of the device. The model is designed to be computationally efficient, highlighting its potential contribution in simulations of next-generation computing architectures.

Keywords: Physical modeling, Compact model, Conductive-Metal-Oxide, TaO_x, HfO_x, Analog ReRAM, Resistive switching, Substoichiometric, Defects

Acknowledgements

This Master's thesis project was carried out at IBM Research Europe, in the Zurich Laboratory. Definitely the best professional experience of my life so far.
During my time at IBM, I learned a lot both academically and personally. Therefore, first of all, I would like to thank all the members of the Neuromorphic Devices and Systems group at IBM Research. It has been an incredible journey and each of you welcomed me and made me feel important. I'm very happy to have met you and I'm grateful for introducing me to a fantastic world. All this is invaluable.
A very special thanks to Valeria and Donato, my supervisors at IBM. I have never learned so much from someone in such a short time. Your ambition inspires me to always push my limits further. My most sincere gratitude to both of you for helping, supporting, and motivating me on projects that not all students have the luck to participate in. Without you, it would not have been the same.
Furthermore, I would like to thank Dr. Stephan Menzel from the Peter Grünberg Institute Forschungszentrum Jülich. His knowledge was crucial for conducting and improving this study, and it was an honor to discuss with him the right paths of this project to follow.
Additionally, during these six months, I had the fortune to participate to the Memrisys2023 International Conference held at Politecnico di Torino. It was really interesting to understand how experts in this field approach to research. I'm grateful to my thesis supervisor and Prof. Carlo Ricciardi and to the IBM team for giving me the opportunity to attend the conference. It is not every day that master's students get such opportunities, and I'm aware of that.
I would also like to thank the team of the PHASTRAC research project (grantID: 101092096) funded by the European Union, to which I contributed with the results of this thesis. It was a pleasure to participate.
Next, a special thanks to my family who shows me unconditional love and support every day and for every choice I make, often not simple. I love you, thank you for being there whenever I need it.
Last but not least, a heartfelt thanks to all my friends. I rarely remind you how fundamental you are for the balance of my life. I often wish I could be closer to you. I love you like few things in the world, and I don't need anything else when you are with me, physically or not. You don't need to be mentioned one by one. If you are important for me, you know.

Thank you all,
Matteo.

Contents

List of Figures **7**

1 Introduction **13**
1.1 Beyond von Neumann computing paradigm 13
1.2 Memristive-based crossbars as deep learning accelerators 15
1.3 Analog ReRAM for AI accelerator 18
1.4 ReRAM compact modeling status 22

2 Methods **25**
2.1 Electrical characterizations 25
2.1.1 Quasi-static voltage sweep measurement 25
2.1.2 Pulse response characterization 29
2.2 Physical modeling 32
2.2.1 Theoretical fundamentals 33
2.3 Numerical approaches for physical modeling 36
2.3.1 Discretization of ordinary differential equations 37
2.3.2 Newton-Raphson numerical solver for non-linear systems 38

3 Modeling and results **41**
3.1 Compact model 41
3.1.1 Physics of resistive switching dynamics 41
3.1.2 Equivalent circuit 45
3.1.3 Algorithm implementation 49
3.2 Compact model validation 50
3.2.1 Quasi-static voltage sweep model 50
3.2.2 Pulse response model 56

4 Conclusions **65**
4.1 Key findings 65
4.2 Future perspectives 66

A Appendix - Fabrication process of HfO_x/TaO_x ReRAM device **69**

B Appendix - Switching time characterizations **73**

C Appendix - Matrix formalism to solve the non-linear system **75**

Bibliography **77**

List of publications **83**

List of Figures

1.1 (a) Classical von Neumann architecture. Adapted from [2]. (b) Need of new computing solutions to overcome von Neumann architecture issues [3]. 13
1.2 Evolution of computing architectures to tackle von Neumann bottleneck [3]. . 14
1.3 Schematic representation of processing unit and conventional (above) or computational (below) memory. Adapted from [11]. 15
1.4 (a) Fundamental two-terminal circuit elements: resistor, capacitor, inductor and memristor [13]. (b) Typical pinched hysteretic I-V characteristic of bipolar memristive devices. 16
1.5 Schematic illustration of crossbar array in 1Transistor1Resistor (1T1R) configuration [15]. 17
1.6 Example of a fully connected neural network with interlayer computations implemented as Matrix-Vector Multiplications in crossbar arrays. 17
1.7 Emerging memristive techonologies [20]. 18
1.8 Anion- or cation-based Resistive Random Access Memory OFF/ON states schematic illustrations [21]. 20
1.9 Baseline Metal-Insulator-Metal filamentary ReRAM device operation. 20
1.10 (a) Comparison of $I-V$ sweep characteristics for monolayer baseline TiN-HfO_2-Ti ReRAM (left) and bilayer TiN-HfO_x-TaO_x-TiN ReRAM (right), adapted from [26]. (b) Comparison of bidirectional accumulative response characteristics for monolayer baseline TiN-HfO_2-Ti ReRAM (left) and bilayer TiN-HfO_x-TaO_x-TiN ReRAM (right), adapted from [27]. 21

2.1 Picture of the chip held on the probe station chuck during the measurement. The tip is in contact with the metallic pad of the DUT and the SMU signal of the voltage ramp is applied from the top contact. 25
2.2 Illustration of the experimental setup employed for the quasi-static I-V sweep characterizations. 26
2.3 Voltage ramp scheme adopted by the parameter analyzer for the quasi-static I-V sweep characterizations. 27
2.4 (a) Forming procedure 1st step: negative sweep. (b) Forming procedure 2nd step: positive forming. (c) Forming procedure 3rd step: first SET. (d) 10 cycles of SET/RESET quasi-static I-V sweep [39]. 28
2.5 10 cycles of quasi-static clockwise I-V sweep [39]. 28
2.6 Illustration of the experimental setup employed for pulse response characterizations. 29
2.7 Pulse sequence READ-READ-Programming SET pulse-READ-READ employed in the switching time experiment. 30
2.8 (a) $I(t)$ evolution as a response to the experimental SET pulse sequence with zoom on the switching transition in the left inset. (b) Superimposed $I(t)$ curves for increasing $|V_{pulse}^{SET}|$. 31

2.9 Experimental pulse scheme to characterize the bipolar accumulative responce of the ReRAM device. 31
2.10 (a) Experimental accumulative conductance response of the device stimulated by 10 batches of pulse streams with 200 up and 200 down pulses. (b) Zoom on 1 central batch to highlight the analog potentiation/depression of the device conductance. 32
2.11 Schematic illustration of ion hopping process within the lattice potential landscape in absence (a) and in presence (b) of an applied electric field. Redrawn from [49]. 33
2.12 Schematic illustration of trap-assisted conduction mechanisms. (a) Trap-to-trap tunneling and (b) Poole-Frenkel emission. Redrawn from [39]. 35
2.13 RC circuit in the thermal domain associated to the Newton's cooling law with Joule heating as external heat contribution. 36

3.1 Classification of point defects in crystalline structures. 42
3.2 (a) Realistic interpretation of multiple oxygen vacancies filaments formation as a consequence of the electroforming step. (b) Unique filament approximation. 43
3.3 Oxygen vacancies spatial arrangement interpretation in CMO/HfO_x ReRAM in LRS (a) and HRS (b). 43
3.4 (a) Direction of the electronic conduction/ion migration. (b),(c) Sketches of spectral and spatial location of defects trap states in the band diagram for LRS and HRS respectively, redrawn from [39]. 44
3.5 Equivalent circuit for the electrical model of CMO/HfO_x ReRAM with TaO_x as CMO. 45
3.6 Flowchart for quasi-static I-V sweep simulation. 50
3.7 Flowchart for pulse response simulation. 50
3.8 Voltage-time characteristic employed to simulate the I-V sweep. 51
3.9 Measured and simulated clockwise I-V characteristics of TaO_x/HfO_x-based ReRAM device in linear (a) and logarithmic (b) scales. 51
3.10 (a) R-V characteristic. (b) Time evolution of oxygen vacancy concentration and ionic migration energy barrier. (c) Time evolution of the average temperature in the dome. (d) Ionic current curing the SET/RESET sweep. 52
3.11 Ionic current drift-diffusion components during SET (a) and RESET (b) switching phases. 53
3.12 Impact of thermal model parameters on the I-V characteristic: (a) C_{th} and (b) R_{th}. 54
3.13 Simulated programming scheme to access IRSs with quasi-static multiple sweeps. 55
3.14 (a) Simulated clockwise I-V characteristic of TaO_x/HfO_x-based ReRAM device for 8 IRSs. (b) Evolution of oxygen vacancies concentration and total resistance of the device over the time interval of multiple sweeps. 55
3.15 Oxygen vacancies spatial arrangement interpretation in TaO_x/HfOx ReRAM in LRS/IRS/HRS. 56
3.16 (a) Simulated pulse sequence READ-Programming SET pulse-READ. (b) $I(t)$ evolution as a response to the SET pulse sequence. (c) Resistance plotted during the time intervals of pre- and post-pulse READ. (d) Temperature evolution during the time interval of the SET pulse sequence. 57
3.17 (a) Simulated pulse sequence READ-Programming RESET pulse-READ. (b) $I(t)$ evolution as a response to the RESET pulse sequence. (c) Resistance plotted during the time intervals of pre- and post-pulse READ. (d) Temperature evolution during the time interval of the RESET pulse sequence. 58

3.18 (a) Temperature evolution during the time interval of an ultra-short RESET pulse. (b) Simulated ultra-short pulse sequence READ-Programming RESET pulse-READ. 58
3.19 SET (a) and RESET (b) conditions to consider the pulse features as data for the Voltage-Time Trade-Off plot. 59
3.20 Experimental Voltage-Time Trade-Off plot showing the exponential $t_{SET}(V_{pulse}^{SET})$ relation in logarithmic scale. 60
3.21 SET Voltage-Time Trade-Off plots with simulated and experimental data in logarithmic (a) and linear (b) scale. 60
3.22 Simulated SET (a) and RESET (b) Voltage-Time Trade-Off plot data fitted with exponential laws to extract the analytic $t_{pulse}(V_{pulse})$ relation. 61
3.23 (a) Simulated programming scheme with 10 up-10 down pulses and 20 READ pulses. (b) $I(t)$ characteristic during the pulse stream time span. (c) Accumulative oxygen vacancy modulation. (d) Bidirectional potentiation/depression characteristic. 62
3.24 (a) 10 cycles of experimental potentiation/depression characteristic. (b) Comparison between simulated and experimental accumulative response using the same programming pulse scheme. 63
3.25 Simulated accumulative response corrected empirically to match the experimental conductance window. 63

A.1 Fabrication process flow of TiN-TaO_x-HfO_x-TiN ReRAM device in cross-section view: (a) Deposition of ReRAM material stack on a Si substrate. (b) Photoresist lithography patterning to define the cell area. (c) Plasma etching. (d) Si_3N_4 passivation. (e) Photoresist lithography patterning to define the via area. (f) Deposition of W top electrode. Materials geometry is not in scale. . 70
A.2 (a) Scanning Electron Miscroscope image of the ReRAM device stack cross-section [41], highlighting the 200 nm width of the cell. (b) Bright field Scanning Transmission Electron Microscope image of the ReRAM device stack cross-section [42]. 71

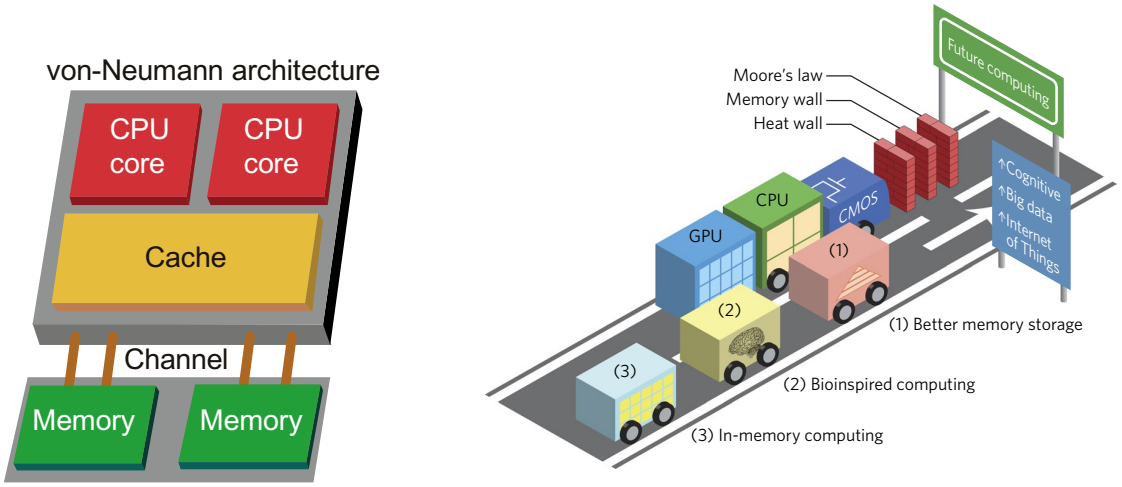
B.1 $I(t)$ evolution in the switching time characterization. 73
B.2 $I(t)$ evolution in the switching time characterization. 73
B.3 $I(t)$ evolution in the switching time characterization. 74
B.4 $I(t)$ evolution in the switching time characterization. 74
B.5 $I(t)$ evolution in the switching time characterization. 74

1 Introduction

1.1 Beyond von Neumann computing paradigm

In the recent years, computing systems encountered several challenges about data processing and storing. This is mainly attributed to the foundational principle of classical von Neumann architectures, i.e. the common ground on which most of modern Information Technology (IT) infrastructures are built on (Fig.1.1a). The von Neumann paradigm [1], following the name of its inventor (John von Neumann), was proposed in 1945 and it is based on the scheme below:

- Central Processing Unit (CPU), divided in control unit and Arithmetic Logic Unit (ALU), is separated by the memory unit.
- A single memory unit is shared to store instruction and data.
- CPU and memory unit communicate through a unique channel (bus), first by sending the instructions and then processing the information.



(a) (b)

Figure 1.1: (a) Classical von Neumann architecture. Adapted from [2]. (b) Need of new computing solutions to overcome von Neumann architecture issues [3].

Although numerous improvements made this architecture extremely mature over the past years, it is inherently limited in terms of power consumption and computational speed. Since data and processing instructions share the same communication path (bus), most of the time is spent for the memory access, i.e. to recall or move data [4]. The overall consequence

is an intrinsic speed limit, which in literature is usually named "*von Neumann bottleneck*". Furthermore, with the scaling of electronic components in microprocessors, the performances of CPUs and memory units followed asynchronous improvement trends: CPUs and memory units performances improved by 50% and 7% per year respectively [4]. As a result, today's CPUs operate in GHz range, while conventional memories bandwidht is of the order of MHz (Memory wall in Fig.1.1b).
As predicted by Moore in 1975, the number of components into silicon Integrated Circtuits (ICs) almost followed the same trend, doubling approximately every 2 years [5]: this was the consequence of a progressive dimensions reduction of logic building blocks, based on Complementary Metal-Oxide-Semiconductor (CMOS) technology. Apart from the scaling issues faced in the last decades (Moore's wall in Fig.1.1b), the operating frequency of Field-Effect Transistor (FET) -based ICs doubled every 2 years [6], leading to a monotonic increase of the power dissipated by microprocessors. Power dissipation problem was already envisioned by Moore: shrinking the processor dimensions led to operate at higher frequencies but the compromise concerned heat dissipation, with limited (and further reduced over the time) surface available for cooling [5] (Heat wall in Fig.1.1b).
Despite the speed improvements allowed to bypass the von Neumann bottleneck and the heat problem was not approaching critical limits, modern infrastructures need new solutions to fulfill the power demand of future computing. Undoubtedly, over the last years, the development of data-centric applications is suffering the lack of appropriate computing systems. For instance, with the growth of Internet of Things (IoT) applications, billions of devices are simultaneously communicating, implying enormous energy consumptions that can exponentially increase in the future [7].
Before 2010, computing operations in Artificial Neural Networks (ANNs) or Deep Neural Networks (DNNs) were designated to architectures based on von Neumann model (CPU computations), while today Graphics Processing Units (GPUs) are employed. The reason rests on the stucture of these architectures, composed by smaller and multiple computing cores (Fig.1.2): GPUs can perform multiple computations in parallel, hence accelerating the computing time. With the relocation of Artificial Intelligence (AI) tasks to GPU systems, the computing performances increased much more rapidly in time (doubling every 3.5 months) [8]. Unfortunately, this approach consumes much more energy, confining data processing only to massive machines. The workloads of latest AIs implemented in the current hardware require computing performances exceeding 10^{18} FLoating point Operations Per Second (FLOPS). The power demand that follows is doubling every 2 months [8].

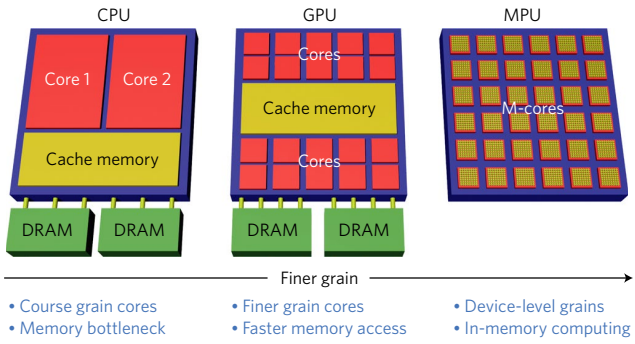


Figure 1.2: Evolution of computing architectures to tackle von Neumann bottleneck [3].

Several low-power strategies have been developed to mitigate the von Neumann bottleneck, such as Near-Memory and In-Memory Computing (NMC, IMC). Nevertheless, NMC is far to be the ultime solution to tackle the hungry of power and speed [9, 10]: moving

computation closer to the data significantly reduces the data transfer time but there is still a separation between processing and memory units [11].
IMC represents the most promising approach, based on having memory elements that can directly perform computational tasks [11]: these Memory Processing Units (MPUs) execute logic operations within memory arrays substituting the role of ALU in von Neumann architectures, as schematically shown in Fig.1.3. MPUs can involve either standard charge-based memory technologies or emerging memory technologies, like resistance-based memory devices. However, charge-based elements such as Static Random Access Memory (SRAM), Dynamic RAM (DRAM) and flash memory, have some drawbacks that impact on the performances of IMC systems: SRAMs and DRAMs are volatile devices, so they need power for data retention, whereas flash memories consume significant power during write and erase operations. Hereby, the most energy efficient solution to develop IMC frameworks consists in including resistance-based non-volatile memories (also known as "*memristors*") as basic hardware elements.

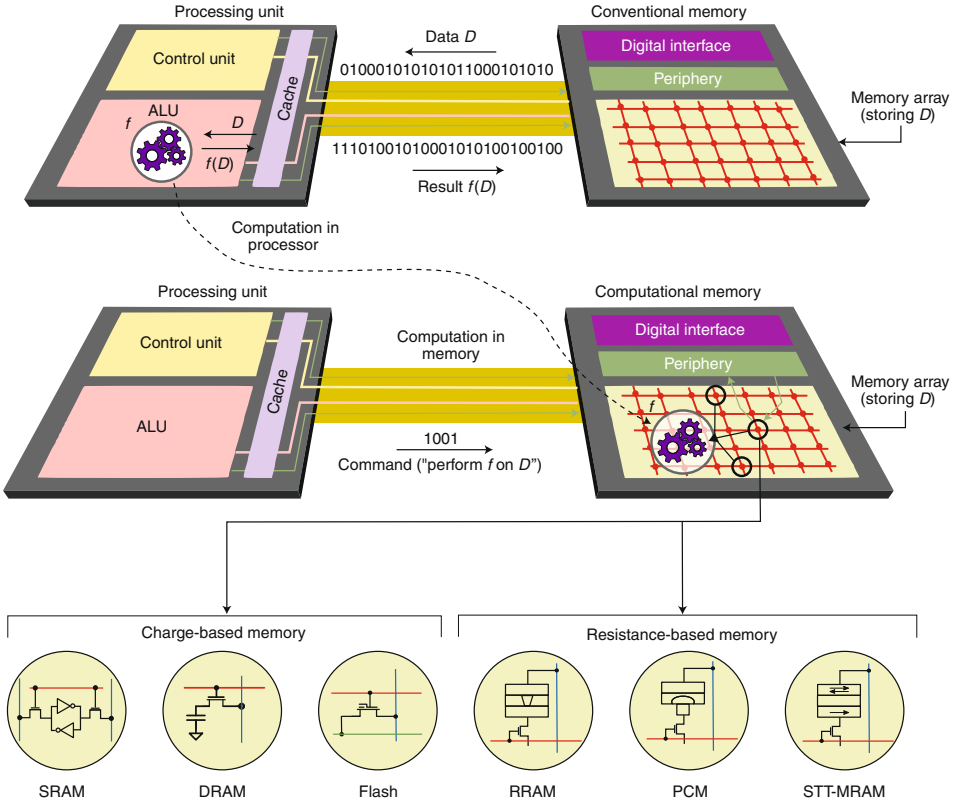
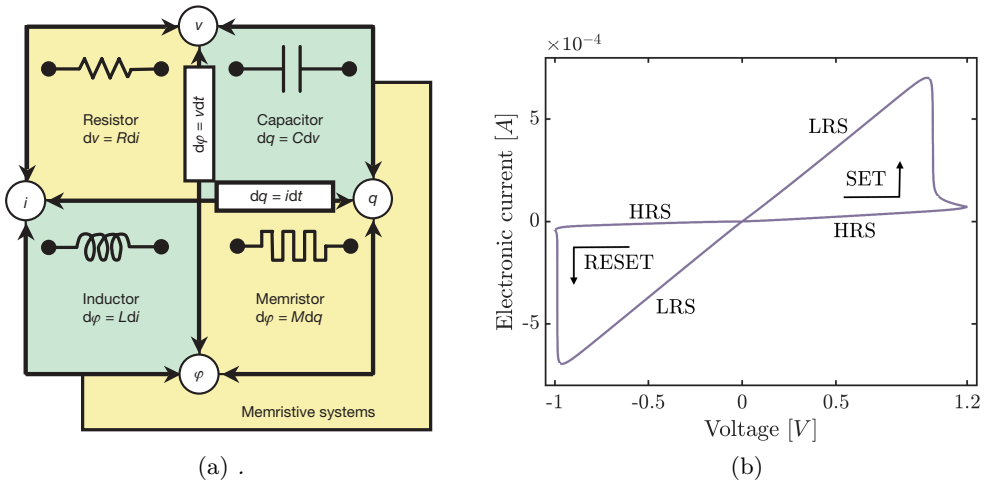


Figure 1.3: Schematic representation of processing unit and conventional (above) or computational (below) memory. Adapted from [11].

1.2 Memristive-based crossbars as deep learning accelerators

Memristive elements were theorized by Chua in '70s [12], motivated by the lack of the fourth basic component to respect the symmetry between electrical independent variables (Fig.1.4a). In circuit elements theory, memristors (as contracted form of "*memory-resistor*") are passive components whose fundamental property is to have a pinched hystertic current-voltage (I-V) characteristic: this means that the resistance depends on the voltage and current history of the device. Moreover, when the device has no power supply, it retains (apart from non-idealities of

the device) the last configuration because of the pinched characteristic (non-volatile memory).



(a) . (b)

Figure 1.4: (a) Fundamental two-terminal circuit elements: resistor, capacitor, inductor and memristor [13]. (b) Typical pinched hysteretic I-V characteristic of bipolar memristive devices.

The operating principle of non-volatile memristors relies on the concept of resistive switching: by applying external electrical stimuli, the resistance of the memristor can be switched between High- and Low Resitance State (HRS and LRS), analogous to the storage of "0" and "1" as digital information. The switching operation from HRS to LRS is called "*SET*". On the contrary, from LRS to HRS it takes the name of "*RESET*". The memristor is unipolar if the voltage polarity required to SET and RESET the device is the same, otherwise it is bipolar. An example of I-V characteristic for bipolar memristor is shown in Fig.1.4b.
Memristors have generated significant interest for the development of next-generation memories and new-computing paradigms due to additional reasons:

- Reading and writing procedures are very fast if compared to standard memory technologies.
- The simple 2-terminal structure entails the chance of large scalability for densely packed memory arrays.

Conventional CMOS technology is based on 3-terminal devices and their size scales as $6F^2$, where F is the feature size of the manufactured device structure. Conversely, memristive devices can be integrated in CrossBar Arrays (CBAs) with 1 memory element in each cross-point and their theoretical scaling follows a $4F^2$ proportionality [14], potentially overcoming the scaling limits of CMOS technology. Additionally, future memory applications are predicted to be based on emerging memristive technologies integrated into CBAs, as the numerous options about material choices enable them to be CMOS and Back-End Of the Line (BEOL) compatible [14].
In order to integrate memristors as cross-point memory elements in CBAs, they are fabricated on top the grounded Bit Line (BL) and below the biased Source Line (SL). The whole memory cell is actually composed of the memory element and a selector biased by a Word Line (WL), which is used to open or close the circuit between SL and BL. The selector can be a FET, a diode or another resistive memory in complementary configuration. Fig.1.5 shows a schematic illustration of a memristor-based CBA and its components.

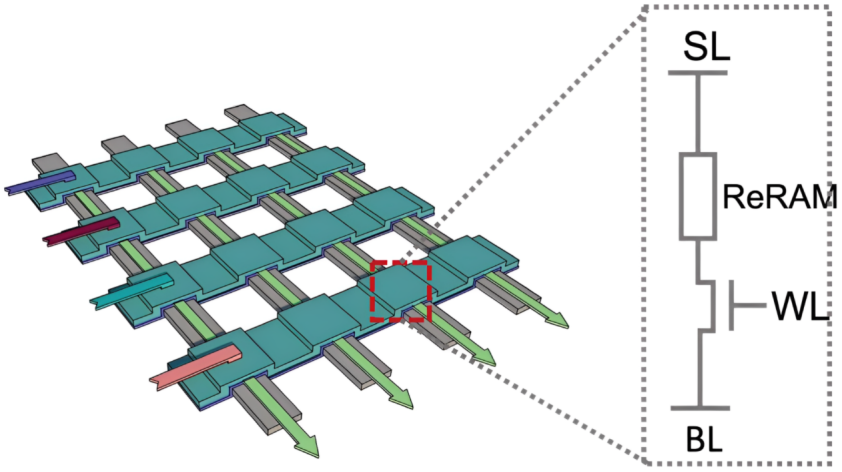


Figure 1.5: Schematic illustration of crossbar array in 1Transistor1Resistor (1T1R) configuration [15].

Cross-point architectures with non-volatile resistive memories represent a relevant breakthrough in the field of brain-like computing system. It has been demonstrated that Resistive Processing Units (RPUs) in cross-point configuration offer a power-efficient hardware solution to perform transmission of data among neural networks layers [16]. In ANNs (or DNNs), many operations involve Matrix-Vector Multiplication (MVM) with parallelized computations that are processed simultaneously. Since CBA RPUs are inherently parallel, they allow massive low-power operations simultaneously (mitigating the energy problem of GPU-based digital accelerators).

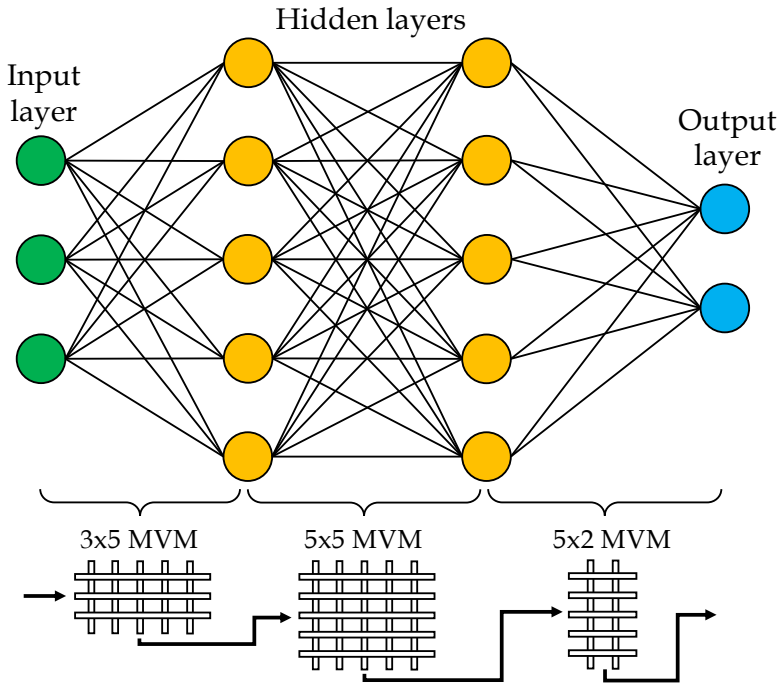


Figure 1.6: Example of a fully connected neural network with interlayer computations implemented as Matrix-Vector Multiplications in crossbar arrays.

In the context of neural networks, the weights of a fully connected layer can be mapped as

a matrix of conductances (associated to the resistive states of memristors) in the CBA, where each row stores the weight of output neurons with respect to the inputs layer. Therefore by applying input voltages to SLs (rows), the current flowing into columns (measured in the BL) represents the output of the fully connected layer, computed as weighted sum of inputs [17]. Mathematically, a fully connected layer in a neural network (Fig.1.6) is represented by the following problem:

$$\bar{\mathcal{V}} \times \mathcal{G} = \bar{\mathcal{I}}$$

where $\bar{\mathcal{V}}$ is the input voltages vector applied to the SLs (rows), $\mathcal{G}$ is the matrix of conductances in the CBA and $\bar{\mathcal{I}}$ is the output currents vector whose components refer to each BL. The synaptic weights are stored in the CBA by programming each memristor. Accordingly, the output current in each column is the results of the weighted sum of as many Ohm's laws as the rows are.

$$I_j = \sum_i V_i \cdot G_{i,j}$$

MVMs performed in this parallel way, are more computationally efficient with respect to the equivalent operations digitally implemented: in the former case, a one shot computation is required to map the output of the neural network layer (key advantage for large networks, scaling as O(1) complexity), while the latter implementation has complexity that exponentially increases for large matrices dimension (O(n) complexity, where n is the matrices dimension) [18].
In summary, hardware implementations based on CBA RPUs represents a revolutionary solutions to address scalability, power demand and parallel computations in novel AI tasks. In addition, CBAs can be composed of analog memory elements, such as multi-state memristor with multiple conductance levels. This can be a crucial requirements both for neural network applications (mapping analog weights during training) and for data integration (storage purposes) [19].

1.3 Analog ReRAM for AI accelerator

There are several emerging memristive technologies that can be employed as novel memory. Fig.1.7 shows the most promising non-volatile memory solutions to replace classical RAMs, all of them based on the same principle of associating the "0"/"1" information to HRS/LRS.

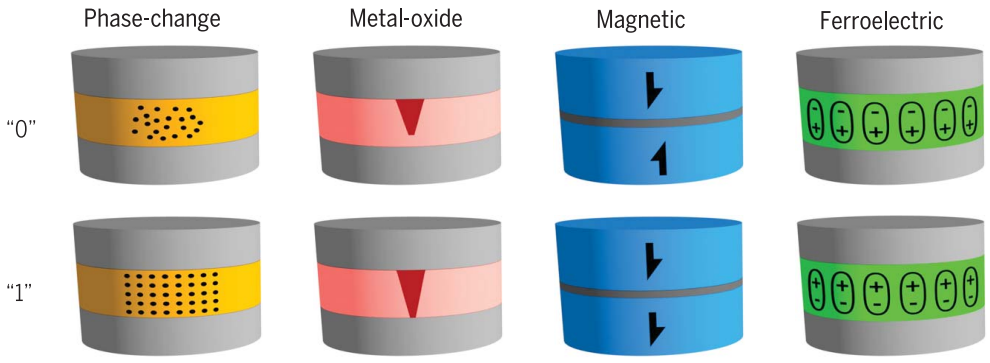


Figure 1.7: Emerging memristive techonologies [20].

Phase-Change RAM (PCRAM) Memristors based on phase-change concept are composed of a chalcogenide (Ge–Sb–Te compound) insulator enclosed between a metallic heater and a metallic electrode. PCRAM switching characteristic is unipolar and the transition among resistive states (SET or RESET) is the consequence of crystallization or amorphization of the chalocogenide material. By forcing a large current for short time, the local temperature in the crystalline (LRS) phase-change material exceeds the melting temperature and the passive thermal dissipation leads to a reconfiguration of the insulator in a disordered phase (HRS). Inversely, a lower current in a longer period allows to reach the crystallization temperature, recovering the chalcogenide ordered phase (LRS). [21].

Ferroelectric RAM (FeRAM) The FeRAM basic idea is to exploit the polarization of a ferroelectric layer, sandwiched between 2 metal layers. The bit is encoded as different alignments of the domain polarization, which can be switched with bipolar electrical stimuli. [21].

Magnetic RAM (MRAM) Magnetoresitive tunnel effect is the fundational principles of magnetic memristors, made of a tunneling oxide (barrier) between 2 magnetic layers. The alignment of magnetic domains (out-of-plane in modern MRAMs) in the 2 magnetic layers can be anti-parallel (HRS) or parallel (LRS). One of the 2 ferromagnetic layers has fixed magnetization (pinned layer), while the free layer magnetization is switched from anti-parallel to parallel (SET) and the other way round (RESET) through opposite current flows. [21].

Resistive RAM (ReRAM) In ReRAMs, data are encoded as different resistive states of a dielectric layer enclosed between 2 metal electrodes. An external voltage applied to the electrodes allows the movement of ionic species (including defects) within the dielectric. Thus the general resistive switching mechanism is attributed to field-, temperature-driven ion migration and chemical potential-driven redox reactions. There are different types of ReRAMs depending on the nature of the migrating ionic species (anion or cation, i.e. oxygen or metal ion) through the lattice of the metal-oxide active layer (Fig.1.8). Since the migrating species are electrically charged, they moves in opposite directions depending on the external stimulus polarity, so ReRAMs are typically non-volatile bipolar devices. The switching process relies on the formation of local conductive paths that allow the current flow between the electrodes (LRS). Their rupture limits the conduction due to physical insulation (HRS). Furthermore, some ReRAM devices exhibit multilevel capabilities, i.e. the possibility to access Intermediate Resistive States (IRSs). [22]

ReRAMs involving oxygen ions migration showed great potential in terms of scalability and CMOS compatibility of their fabrication processes: regarding these two aspects, filamentary ReRAM are the most advanced with respect to area-type ones [23]. For these reasons, the focus of this work is on filamentary ReRAMs, so a more detailed description of the technology is provided in the following.

As mentioned above, in ReRAMs the alteration of the conductive properties of a metal-oxide causes resistive switching phenomena. The choise of the active material are various, including binary/ternary metal-oxide: the most common materials are CuO_x, WO_x, HfO_x, TiO_x, TaO_x, $SrTiO_x$ [21]. Conventional materials as metal electrodes are Ti, Cu, TiN, Pt, W [24].
In the pristine state after fabrication, the Metal-Insulator-Metal (MIM) structure is insulating, so a configuration step called "*electroforming*" (or simply "*forming*") is necessary to generate a soft and partial breakdown in the dielectric [22]. This holds in general, as the

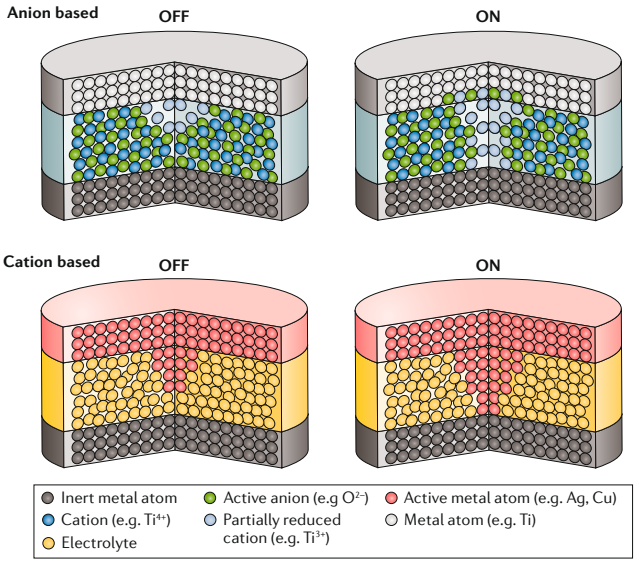


Figure 1.8: Anion- or cation-based Resistive Random Access Memory OFF/ON states schematic illustrations [21].

forming step is not always required (forming-free ReRAMs). After the forming phase, the resistance of the shunted dielectric layer can be controlled by applying external biases, either in the form of triangular voltage sweeps or square voltage pulses.
ReRAM devices can be classified as area-type and filamentary. In area-type ReRAMs, the migration of ions occurs uniformly across the entire active layer cross-section. On the other hand, in filamentary ReRAMs, the conductive path is unique, also referred to as "*conductive filament*". Typically, binary oxide-based ReRAMs are filamentary. Electrical characterizations of HRS/LRS values as a function of the cell area allows to demonstrate if the switching is either filamentary or area-dependent [25]: when HRS/LRS are cell size-independent the device is demonstrated to be filamentary-type.
In filamentary ReRAM devices, during the SET transition (HRS $\rightarrow$ LRS) ions migrate until the conductive filament bridges the metal electrodes. Whereas, during the RESET phase (LRS $\rightarrow$ HRS) ions migrate in the opposite direction up to the formation of an insulating gap between the remnant conductive filament and the electrode (Fig.1.9) [22]. The reversible formation and rupture of localized conductive filaments within the dielectric layer results in abrupt changes of resistance. Filamentary ReRAM devices typically suffer from variability of resistive states, stochasticity and noise [22]: the stability of data is affected as a consequence.

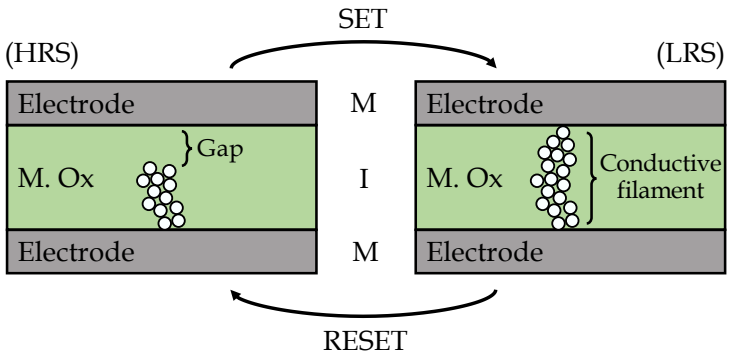
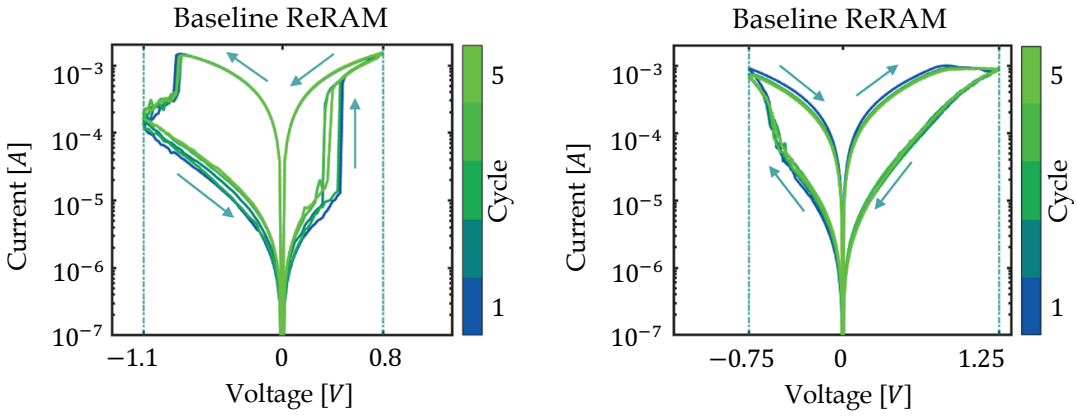


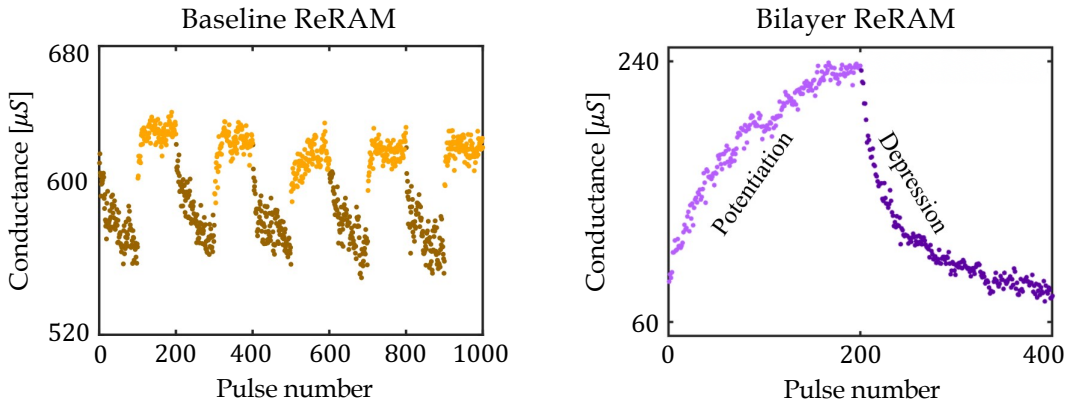
Figure 1.9: Baseline Metal-Insulator-Metal filamentary ReRAM device operation.

The drawbacks of filamentary ReRAM devices based on baseline MIM stack can be mitigated through material level improvements: it has been demonstrated that integrating another substoichiometric metal-oxide layer between the dielectric and the electrode can significantly improve the device performances. Following these considerations, the Neuromorphic Devices and System group at IBM Research Europe optimizeded the switching characteristics of monolayer baseline TiN-HfO_2-Ti ReRAMs by transitioning to bilayer TiN-HfO_x-TaO_x-TiN ReRAMs [26]. In particular, the bilayer approach allowed for:

- Improved symmetry between SET/RESET characteristics.
- Relaxed abruptness of switching processes.
- Reduce stochasticity of SET/RESET cycling.



(a)



(b)

Figure 1.10: (a) Comparison of $I - V$ sweep characteristics for monolayer baseline TiN-HfO_2-Ti ReRAM (left) and bilayer TiN-HfO_x-TaO_x-TiN ReRAM (right), adapted from [26]. (b) Comparison of bidirectional accumulative response characteristics for monolayer baseline TiN-HfO_2-Ti ReRAM (left) and bilayer TiN-HfO_x-TaO_x-TiN ReRAM (right), adapted from [27].

Achieving symmetric SET/RESET switching operation is a crucial requirement for implementing ReRAM devices in RPUs for DNN training applications [16]. Therefore, bilayer ReRAMs, which exhibit reduced asymmetry compared to baseline ones, are better suited for this hardware application. Furthermore, as reported by Wu et al. [28], introducing a properly designed metal-oxide in bilayer ReRAMs results in more gradual SET/RESET transitions:

this is attributed to an improved heat confinement control. In monolayer baseline ReRAMs, the lack of temperature control leads to self-accelerated SET processes [29], causing abrupt transitions. Gradual resistive switching characteristics are essential for enabling analog weight updates in RPUs, as an improved graduality relfects a larger number of conductance states (IRSs) accessed during NN hardware demonstrations [30]. As shown in Fig.1.10, both the I-V sweep (Fig.1.10a) and the conductance accumulative response (Fig.1.10b) of bilayer devices are characterize by symmetric and gradual switching operations and improved cycle-to-cycle stochasticity.
The bilayer ReRAM approach, rather than a monolayer baseline device, offers improved performances and reliability of the device (if well engineered) but the structure is more complex from a fabrication standpoint.

ReRAMs have gained significant attention as storage devices due to their superior properties with respect to standard memory technologies, such as the low-power consumption, non-volatility, multi-state capacity, fast write/read operations, higher density and potential large scalability due to the involved nanoscale phenomena [31]. Especially in the neuromorphic computing field, ReRAM devices represent a promising hardware implementation solution to perform both binary IMC [32] and Analog In-Memory Computing (AIMC) [33, 34] within CBA architectures. The energy efficiency of parallel weight updates in CBA with analog ReRAMs makes these devices attractive to develop deep learning accelerators for training applications, offering outstanding speed and power consumption improvements compared to CPU/GPU counterparts. BEOL-integrated ReRAM CBA deep learning training has been already demonstrated by IBM with their analog bilayer technology [30].

1.4 ReRAM compact modeling status

In order to accelerate the circuit-level integration of ReRAM devices for IMC and neuromorphic applications, simulation models are urgently required.
The hierarchy of simulation models is wide and each of them is suited for particular purposes. For instance, ab initio approaches allows to simulate material fundamental properties and Finite Element Modeling (FEM) can be used to address electro-thermal simulations. Only semi-empirical or compact models are suited to describe the device behavior at the circuit-level [35]: these models run in short timescales, so the result is a trade-off between accuracy and efficiency of the simulation. However, an accurate description of the device can be achieved by developing physics-aware compact models, without fitting parameters and purely mathematical equations that lead to a loss of model predictivity [36].
The design of large circuits involving ReRAMs is crucial for pre-fabrication verifications, new applications proof of concepts and identifications of potential circuit optimization. Typical IC design tools that enable the simulation of ReRAM-based circuits need robust models for the single device characteristics. The model must be able to capture the main electrical features, such as I-V characteristic, resistive switching dynamics, failure mechanisms (endurance, retention) and variability. Hereby the core of this work concerns the development of a fully-dynamic[1] physics-aware compact model to describe analog resistive switching operations in filamentary ReRAMs based on conductive-metal-oxide/HfO_x bilayer IBM technology. Other physics-based compact models of baseline HfO_x ReRAM devices has been already proposed, while a tailored one for the ReRAM technology studied in this work is still missing. Therefore, chapter 3 will be focused on the development of a customized model by leveraging existing works found in literature [37, 38].

[1] "*fully-dynamic*" means that the model accounts relevant dynamic aspects of the device, capturing its real time-dependent behavior.

Parameter extraction is a crucial procedure in ReRAM compact models, especially when the model backbone involves physics equation: all the parameters should be physically reasonable as well, such as to be consistent with the "*physics-based*" label. It is common to use kinetic Monte-Carlo FEM models for parameters calibrations [35]. Accordingly, existing FEM models for the same device of this work [39, 40] has been used as groundworks to establish the physically plausible ranges of some parameters used in the simulations.

2 Methods

The following chapter aims to analyze the experimental and theoretical methodologies supporting the results of this dissertation.
In particular, the experimental part is focused on the electrical characterizations of the ReRAM devices, both in quasi-static and AC domain. The obtained characterization data will be compared with simulation ones in the modeling & results chapter (3).
The theoretical sections address the physical mechanisms on which this work relies on and the numerical approaches implemented in the simulations.
The fabrication process flow of bilayer ReRAM device was already established in previous works [26, 41, 42], so it is summarized in appendix A to have a complete overview of the device structure that is electrically characterized and modeled in this project.

2.1 Electrical characterizations

This section is aimed to explain the experimental setups employed for the electrical characterizations of the TiN-TaO_x-HfO_x-TiN ReRAM device.

2.1.1 Quasi-static voltage sweep measurement

Conventionally, the first electrical characteristic extracted for ReRAM devices is the SET/RESET I-V sweep. In TaO_x/HfO_x-based ReRAM case, SET/RESET I-V sweeps are obtained by forcing negative/positive triangular voltage ramps to the device Top Electrode (TE), according to the polarity reported in [26]. To perform these electrical tests, the chip with the fabricated devices is fixed on a SÜSS MicroTec probe station, whose ceramic chuck is equipped by a vacuum pump system to ensure a good contact and to secure the chip position. The chuck is coated with gold and it is set to ground (GND), such as to have the bottom contact of the devices to GND as well. Additionally, the chuck is provided by a computer interface-controlled stage motor to move the chuck (and the chip too) during the experiment. The allowed directions are both the vertical and the horizontal ones.

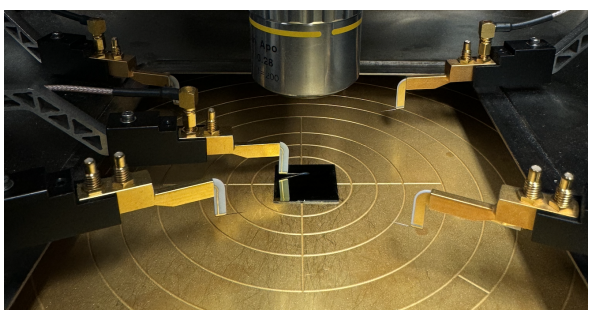


Figure 2.1: Picture of the chip held on the probe station chuck during the measurement. The tip is in contact with the metallic pad of the DUT and the SMU signal of the voltage ramp is applied from the top contact.

An electrically conductive probe microtip (Fig.2.1) is used to contact the TE pad of the RaRAM Device Under Test (DUT) and a Cascade Microtech DPP210 3 axis micromanipulator allows to align the tip and the device pad with high precision.
The electrical signal for the voltage ramp is applied by an Agilent B1500A parameter analyzer with multiple Source Measurement Units (SMUs): each SMU simultaneously forces the voltage and measure the current in the terminal. Furthermore, triaxial cables are used for the electrical connection between the SMUs and:

- the tip, for the applied signal to the TE of the DUT
- the chuck, for the common voltage, i.e. the earth GND.

The setup described so far is schematically represented in Fig.2.2.

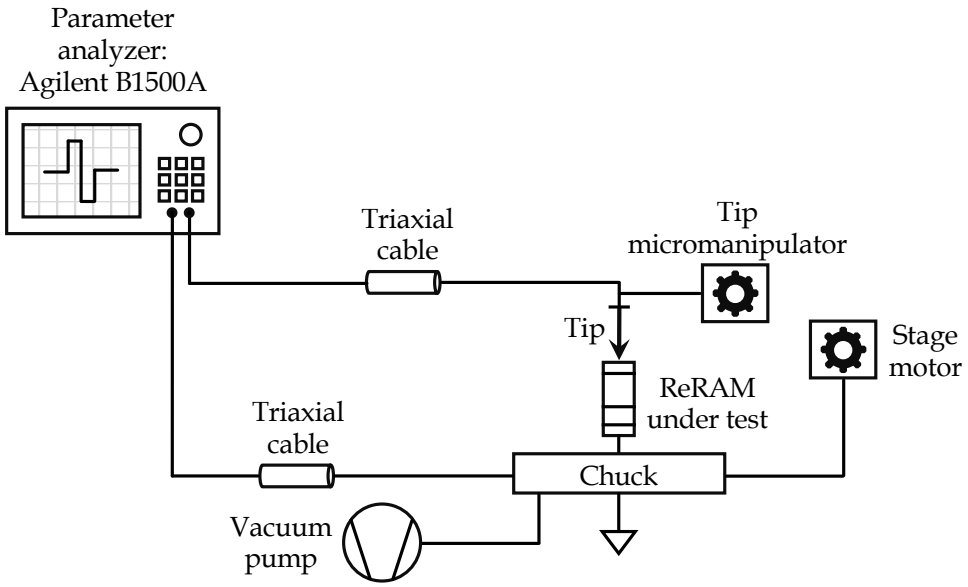


Figure 2.2: Illustration of the experimental setup employed for the quasi-static I-V sweep characterizations.

In order to run a voltage ramp, the parameter analyzer adopts the scheme shown in Fig.2.3. A staircase voltage ramp is the discrete version of the ideal triangular ramp, where the applied voltage is increased for each step by ΔV_S. Each ΔV_S is kept over the time interval corresponding to the hold time Δt_{HT}, during which the current measurement is carried out. The SMU measurement starts only after the delay with respect to the end of the previous voltage step hold time: this time interval is the wait time Δt_{WT}. The Sweep Rate (SR) of the staircase voltage ramp is computed as:

$$\mathrm{SR} = \frac{\Delta V_S}{\Delta t_{HT}}$$

Each voltage sweep includes a forward ramp up to the stop voltage V_{Stop} and a backward one. V_{Stop} is a parameter that can be set for the characterization and different values can be used for SET and RESET sweeps. Moreover the parameter analyzer allows to use a current compliance I_{cc} to limit large currents flowing into the DUT. Despite the availability of this control limit, I_{cc} is never used in the experiments reported in this work.

Before executing quasi-static I-V sweep characterizations, the DUT must go through the forming procedure, which is carried out with the same experimental setup as the one used for SET/RESET cycling: the forming procedure for TaO_x/HfO_x-based ReRAM device is divided

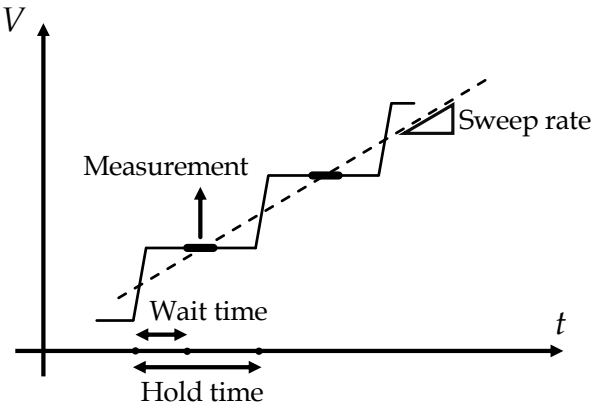


Figure 2.3: Voltage ramp scheme adopted by the parameter analyzer for the quasi-static I-V sweep characterizations.

into 3 steps, which consist in positive and negative sweeps with larger V_{Stop} than the ones involved in the SET/RESET cycling. Positive (or negative) sweep means that $V_A > 0$ (or $V_A < 0$) is applied to the TE.
The forming procedure consists in the following routine:

1. Negative sweep up to $V_{Stop} = -5\ V$. The hysteretic sweep shown in Fig.2.4a is attributed to a further reduction of TaO_x, as the large applied voltage can generate new oxygen vacancies [43] increasing its conductivity.

2. Positive sweep up to $V_{Stop} = +4\ V$ (Fig.2.4b). During this step the device exhibits the typical I-V characteristic associated to a dielectric breakdown [44]. For this reason, this second step is associated to the formation of a conductive filament made of oxygen vacancies in the HfO_x, with forming voltage $V_F = 3.6\ V$.

3. Negative sweep up to $V_{Stop} = -3\ V$ (Fig.2.4c). The device is already formed after the second step, so the aim of this first strong SET (third step) is to bring the device in the 10 $k\Omega$ domain, where it is expected to operate.

In the first step of the forming routine, a 100 $k\Omega$ resistor is used in series with the tip that is in contact with the top metallic pad of the DUT: the series resistor allows to prevent unwanted large currents flowing into the device. In the second and third steps, a 10 $k\Omega$ resistor in series with the tip is used for the same purposes. Series resistors are preferred to limit the current rather than I_{cc}, as they allow to prevent current overshoot [35].
The SR of each forming sweep and SET/RESET cycling is set to SR $= 0.1\ Vs^{-1}$, with $\Delta t_{WT} = 0\ s$ and $\Delta t_{HT} = 100\ ms$: this means that the measurement time coincides with Δt_{HT}. To get SR $= 0.1\ Vs^{-1}$, the ramp has $\Delta V_S = 10\ mV$.

As shown in Fig.2.4d, after the forming procedure the device is ready for SET/RESET cycling. For the SET sweep, the stop voltage is set to $V_{Stop} = -0.9\ V$, while for the RESET, $V_{Stop} = -1.1V$.
A zoom on the 10 cycles of quasi-static I-V sweeps is reported in Fig.2.5, showing an ON/OFF ratio of ~ 3 and negligible cycle-to-cycle variability. In addition, in the clockwise I-V sweeps the current transitions associated to SET and RESET switching processes are more gradual with respect to baseline technologies [26]: typically baseline ReRAM devices made of Metal-Insulator-Metal (MIM) stack exhibit steep SET transitions as a consequence a self-accelerated SET process [29]. Conversely, in TaO_x/HfO_x-based ReRAM device the SET abruptness is reduced, complying with the analog behavior of the device.

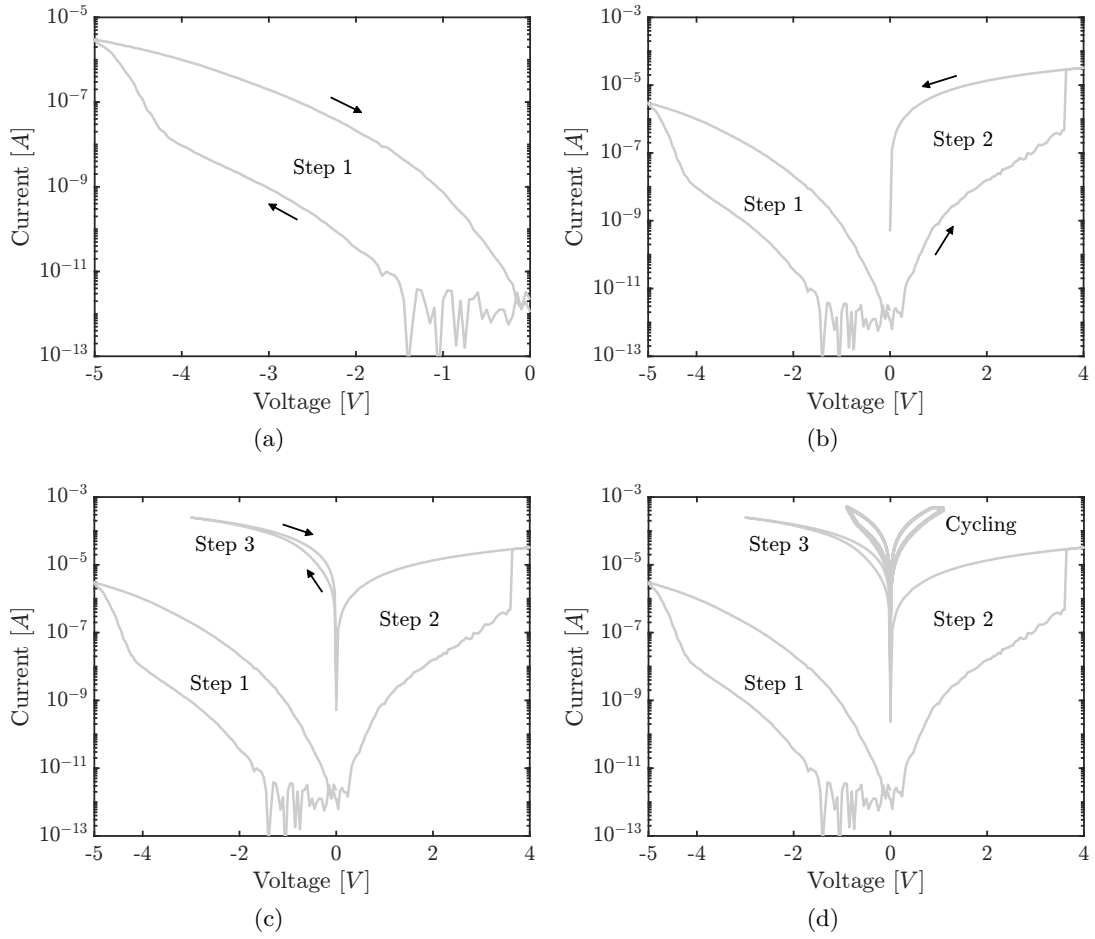


Figure 2.4: (a) Forming procedure 1st step: negative sweep. (b) Forming procedure 2nd step: positive forming. (c) Forming procedure 3rd step: first SET. (d) 10 cycles of SET/RESET quasi-static I-V sweep [39].

The experimental data about quasi-static I-V sweeps (Fig.2.5) will be used as a reference for a first validation of the compact model derived in the next chapter.

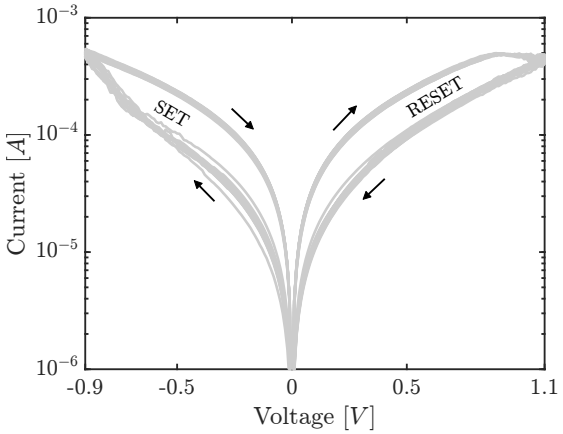


Figure 2.5: 10 cycles of quasi-static clockwise I-V sweep [39].

Nevertheless, the quasi-static I-V sweep is only a proof of concept of resistive switching phenomena [35] because ReRAM devices in real ICs do not operate under these conditions.

2.1.2 Pulse response characterization

A crucial figure of merit for resistive switching devices is the switching time, while another important feature is the conductance accumulative response: to characterize both of them, the application of short ($\sim \mu s \div ns$) voltage pulses [44] is required. The experimental setup (Fig.2.6) employed to apply square voltage pulses to the DUT is described in the following.

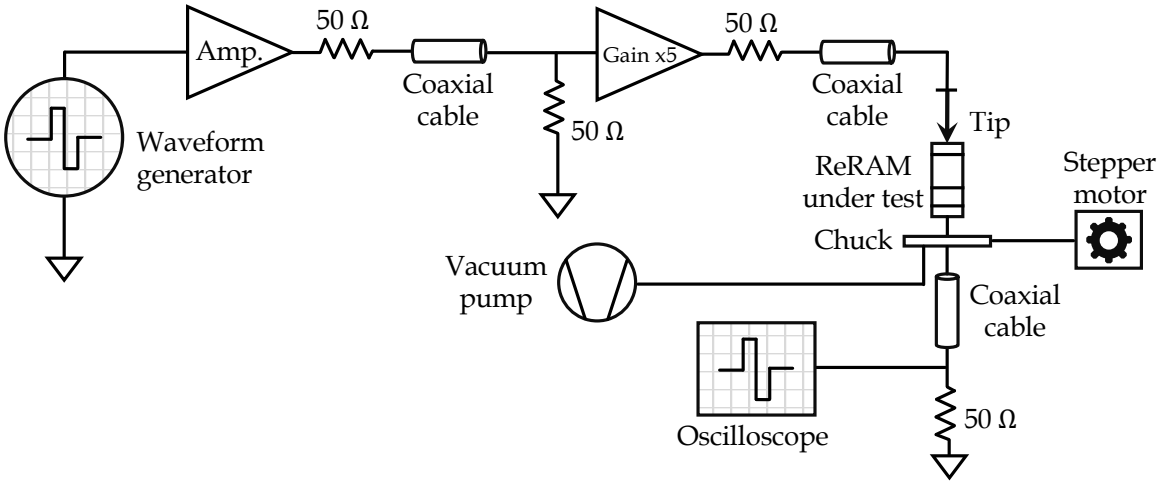


Figure 2.6: Illustration of the experimental setup employed for pulse response characterizations.

The chip with the ReRAM DUT is held on a conductive chuck and its position is secured by a vacuum pump system. Vertical and horizontal displacements of the chuck are controlled by a stepper motor through a LabVIEW software interface: the finest movement allowed is 5 μm in every direction, which is enough to adjust the DUT position with high precision. Contrary to the setup in Fig.2.2, where the tip can be controlled by a micropositioning system, here (Fig.2.6) the tip in contact with the top metallic pad of the DUT is fixed in space. Therefore only the chuck moves to contact the tip with the chip.
A waveform generator (16-Bit 400 MS/s NI PXIe-5451 by National Instruments) assisted by the LabVIEW interface is used to build the arbitrary signal that is first amplified and then applied to the DUT. The input and output impedences of each component of the setup are matched with 50 Ω loads. The generated signal is applied to the TE of the ReRAM DUT through the conductive tip, while the output signal is measured by an oscilloscope (400 MHz NI PXIe-5164 by National Instruments) on a load resistor of 50 Ω in series with the DUT. The load resistor is set to the common voltage of the setup, which in turns is set to the earth GND.
Coaxial cables are used for the electrical connections of the setup. Triaxial cables were not available, although they would fit better the requirements of the experiments carried out with this setup. Triaxial configuration improves the capacitive decoupling of the signal together with low leakage currents: as a result, the RC transients of the setup reduce, so the measured ones can be associated to the device only. In the switching time characterization explained in the following, the measure of the time intervals is affected by the RC transients, hence by employing coaxial cables it is not possible to isolate the device contribution by the setup one.

Switching time characterization

An important requirement in ReRAM device concerns the write operation: Waser et al. [45] discussed about the voltage-time dilemma, i.e. the conditions of the write process to achieve

large retention (up to 10^{16} s). Dealing with square voltage pulses to program devices characterized by non-linear switching mechanisms, the write voltage should be at most ten times larger than the read one. Moreover programming write pulses operating in the ns regime (shorter time intervals are even better) are preferred.
To address the voltage-time dilemma, it is crucial to characterize the non-linearity of the switching time as a function of the writing voltage. In the following, the SET switching time is measured by tuning the amplitude of the programming pulse up to $V_{pulse}^{SET} \lesssim 10 \cdot V_{pulse}^{READ}$, to be consistent with the voltage-time dilemma requirements. $V_{pulse}^{READ} = \pm 0.2\ V$ is the amplitude of the pulse used to read the resistance state of the device, sufficiently small to ensure no switching during the procedures. While V_{pulse}^{SET} is the amplitude of the single square voltage pulse needed to switch from HRS to LRS considering the full resistance window of the device: consistently with the quasi-static operation, HRS = 8 $k\Omega$ and LRS = 2 $k\Omega$ when $V_A = \pm 0.2\ V$.
Fig.2.7 depicts the pulse scheme employed in the switching time experiment, which consists in reading the state of the device before and after the programming pulse, to check that the transition occurs effectively from HRS to LRS. The READ procedure is splitted in a $+0.2\ V$ and $-0.2\ V$ pulses and the resistance is computed as the average between them: this allows to remove potential offsets of measurement setup. The switching time t_{SET} coincides with the width of the programming pulse Δt_{pulse} only if the pre- and post-pulse READ return 8 $k\Omega$ and 2 $k\Omega$ respectively. The experiment is carried out by checking this condition for each V_{pulse}^{SET}. 10 V_{pulse}^{SET} are chosen going from $-1.35\ V$ to $-1.8\ V$ with $-0.05\ V$ steps.
The typical $I(t)$ response of the device associated to a single pulse SET [46] is shown in Fig.2.8a, where the inset shows the absolute increase of the current (SET) during the switching. The $I(t)$ curves for all the programming pulses are reported in the appendix B.
In Fig.2.8b all the $I(t)$ curves associated to the same HRS $\rightarrow$ LRS are superimposed to show that V_{pulse}^{SET} and $\Delta t_{pulse} = t_{SET}$ are inversely proportional (READ sequences are removed from the total curve). However, it is not totally true that $\Delta t_{pulse} = t_{SET}$ experimentally: in Fig.2.8a it is evident that the $I(t)$ response is affected by RC transients, i.e. the current peaks arising when the applied voltage changes. This RC phenomena can be associated to the charging time of the device, which has an intrinsic capacitance [42] or to the limited capabilities of the setup (e.g. biaxial instead of triaxial cables). Since it is hard to decouple the two RC contribution, the time associated to these transients are subtracted from Δt_{pulse} in order to determine t_{SET} used to build the Voltage-Time Trade-Off (VTTO) plot in the results chapter (3). The average RC time considering all the measurements (see appendix B) is estimated to be 280 ns. Additionally, $\tau_{RC,min} = 280\ ns$ determines the measurement limit of this experiment because transients in smaller time scale would not be appreciated.

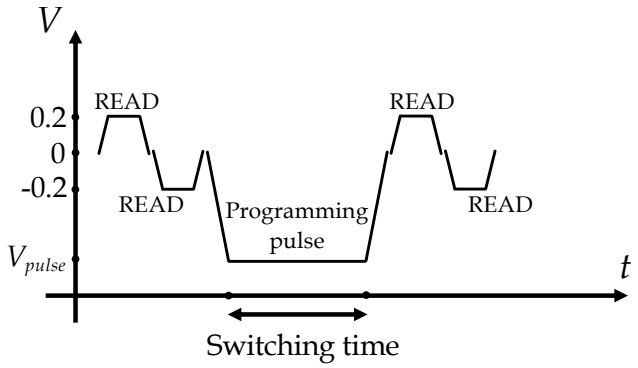


Figure 2.7: Pulse sequence READ-READ-Programming SET pulse-READ-READ employed in the switching time experiment.

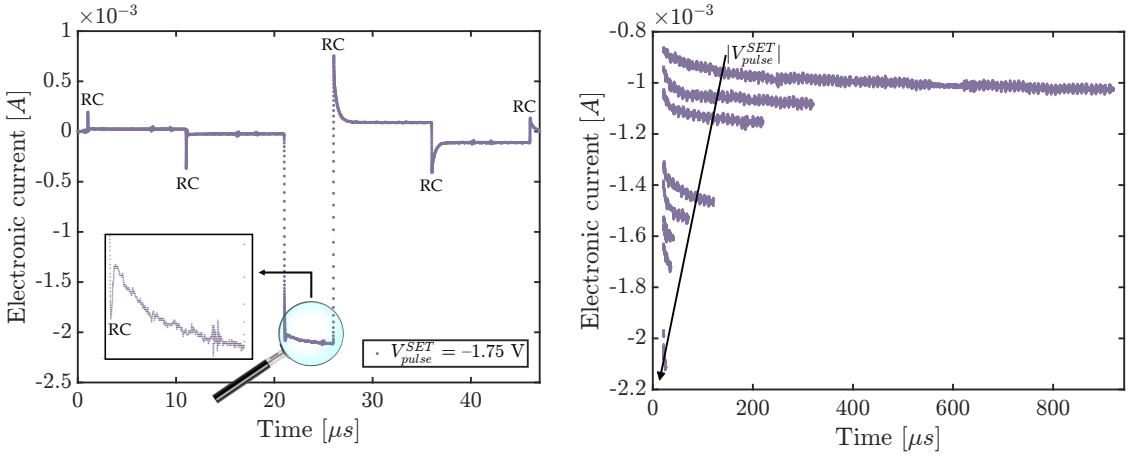


Figure 2.8: (a) $I(t)$ evolution as a response to the experimental SET pulse sequence with zoom on the switching transition in the left inset. (b) Superimposed $I(t)$ curves for increasing $|V_{pulse}^{SET}|$.

It is worth to specify that the switching time is not limited at $\tau_{RC,min}$ by the physical internal processes related to the switching dynamics, while it is limited by the combination of setup limits and the capacitive charging of the cell [46].

Conductance accumulative response characterization

The experimental setup for pulse response characterization (Fig.2.6) is employed to characterize the analog bidirectional switching properties through the accumulative response of the conductance up to a pulse stream. The goal is to characterize the device response when stimulated by short ($\mu s \div ns$) identical voltage pulses. Generally, ReRAM devices undergoing pulsed voltage stress return a resistance (or conductance) update that depends on multiple factors, such as the amplitude/width of the pulse and the resistive state before the update. The pulse scheme adopted in this electrical characterization is schematically represented in Fig.2.9. The experimental pulse scheme consists of a sequence of identical writing pulses with positive polarity $V_A > 0$ (RESET) alternated with READ pulses to measure the conductance state after each programming operation. Then the same sequence is repeated for the negative polarity $V_A < 0$ (SET). The conductance of the device decreases/increases when positive/negative pulses are applied to the TE of the DUT, complying with the polarity reported in previous experiments.

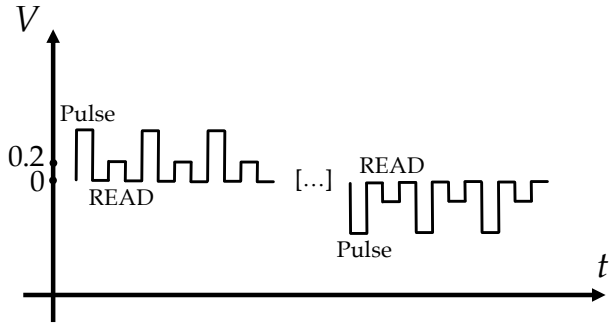


Figure 2.9: Experimental pulse scheme to characterize the bipolar accumulative responce of the ReRAM device.

The DUT undergoes 10 batches of pulse stream, where each batch is composed of 200 positive (up) and 200 negative (down) pulses having the following features:

- $\Delta t_{pulse}^{up} = \Delta t_{pulse}^{down} = 200\ ns$
- $V_{READ} = +0.2\ V$
- $V_{pulse}^{up} = +1.75\ V$
- $V_{pulse}^{down} = -1.25\ V$

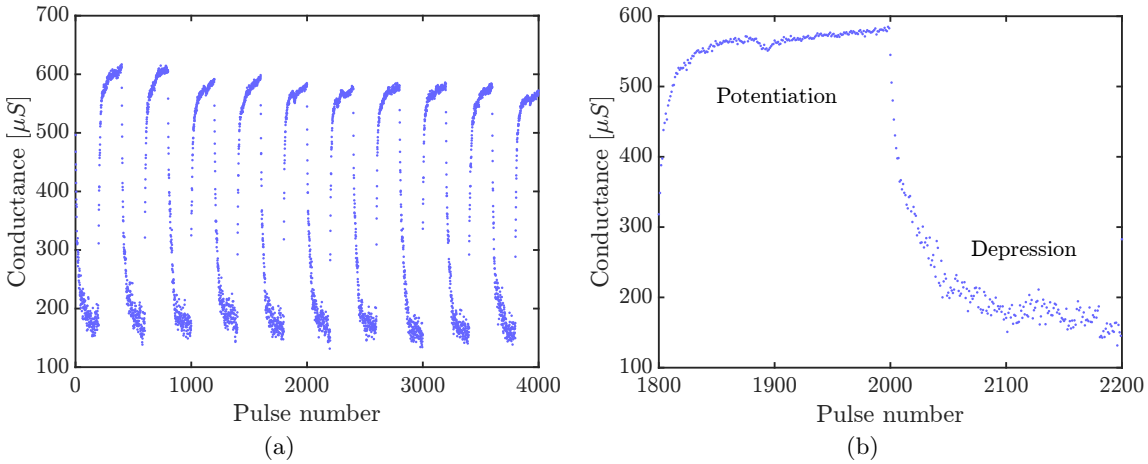


Figure 2.10: (a) Experimental accumulative conductance response of the device stimulated by 10 batches of pulse streams with 200 up and 200 down pulses. (b) Zoom on 1 central batch to highlight the analog potentiation/depression of the device conductance.

Fig.2.10a and Fig.2.10b demonstrate that, even when stressed by trains of short and identical voltage pulses, the ReRAM device exhibits switching properties in analog fashion: the DUT can be programmed in an IRS, as the update of the conductance is incremental with the pulse number in the potentiation (conductance increase) or depression (conductance decrease) curves.
This characterization is aimed to compare the experimental properties of the device with the simulation results of the compact model developed in the next chapter, to understand if the model is able to catch the analog features of the switching mechanisms.

2.2 Physical modeling

Once the experimental data about electrical characterizations has been collected to compare the real behavior of the device with the simulations, it is necessary to establish the basis for the physical modeling of the ReRAM device.
This section focuses on the theoretical fundamentals behind the interpretation of resistive switching processes. Ion migration in solids, electron transport in defective oxide and Joule-heating-based temperature dynamics are the physical mechanisms on which the compact model in this study relies on. The concept is to take advantage of well known equations describing the mechanisms listed above and apply them to model resistive switching processes in filamentary TaO_x/HfO_x ReRAM device.

2.2.1 Theoretical fundamentals

Ion migration

The phenomenon of ion migration in solids can be attributed to multiple physical driving forces such as electric potential gradient (drift), concentration gradient (Fick diffusion) or temperature gradient [47]. Typically the movement of ions is a combination of drift (induced by the electric field) and temperature-enhanced migration.

- The application of an external electric field ($\mathcal{E}$) entails an alteration of the ionic potential landscape and charged ions experience a driving force that might imply their motion in the crystal lattice. Positively charged ions are attracted towards the negative electrode, while negatively charged ions migrate towards the positive electrode.
- Local temperature (T) increase can provide the necessary energy for ions to overcome migration energy barriers enabling their motion within the crystal lattice. At elevated temperatures, the mobility of ions increases, facilitating their migration inside the solid.

Mott & Gurney [48] derived a simple 1D model to describe the combination of these two mechanisms: the atomistic model corresponds to ion hopping between lattice site within the crystal.
In absence of applied electric field, ions can hop to unoccupied lattice sites by overcoming the energy barrier (ΔW_A) in the potential landscape, without any preferential direction, i.e. forward or backward jump (Fig.2.11a). The model takes into account jumps between neighboring lattice sites, so the hopping distance (a) coincides with the lattice constant of the crystal.
In presence of an applied electric field, the hopping energy barrier height lowers (or raises) for forward (or backward) jumps, so the drift of ions exhibits a preferential direction (Fig.2.11b). $|z|q$ denotes the ion charge (where z is the valence of the ion and q the elementary charge) and $\Delta W_A^{f,b}$ is the ion migration barrier for forward ('f') or backward ('b') jump.

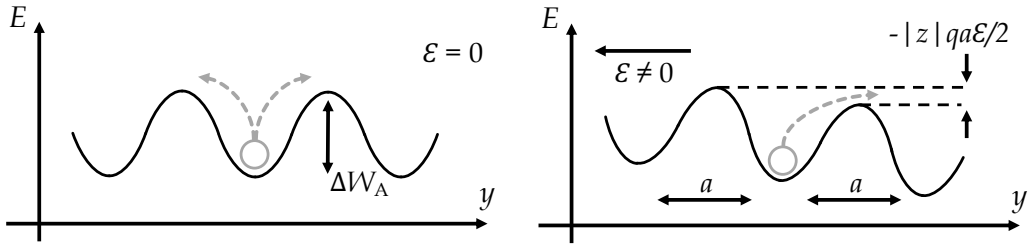


Figure 2.11: Schematic illustration of ion hopping process within the lattice potential landscape in absence (a) and in presence (b) of an applied electric field. Redrawn from [49].

Accordingly, the ion migration barriers modify as:

$$\Delta W_A^{f,b} = \Delta W_A \mp |z|qa\mathcal{E} \cdot \frac{1}{2} \tag{2.1}$$

A non-null drift current is associated to the ion hopping process, considering that forward and backward jump rates follow Boltzmann distributions:

$$J_{ion,drift} = zqC \cdot v_{ion,drift} = zqCa\nu_0 \cdot \left[\exp\left(-\frac{\Delta W_A^f}{k_B T}\right) - \exp\left(-\frac{\Delta W_A^b}{k_B T}\right)\right] \tag{2.2}$$

where C is the concentration of ions and ν_0 is the jump attempt frequency. By inserting equation 2.1 in 2.2, the Mott-Gurney law for ion hopping [48] reads:

$$J_{ion,drift} = zqCa\nu_0 \cdot \exp\left(-\frac{\Delta W_A}{k_B T}\right) \cdot 2\sinh\left(\frac{zq\mathcal{E}a}{2k_B T}\right) \tag{2.3}$$

Equation 2.3 describes the ion drift, with a non-linear dependence between the ion drift velocity ($v_{ion,drift}$) and the applied electric field that comes from the alteration of the energy barriers in the potential landscape. Nevertheless, diffusive driving forces arising from the migration itself are not taken into account by the Mott-Gurney hopping law. Noman et al. [50] modified the original derivation by Mott & Gurney by distinguishing the ion concentration C for left and right sides over the migration energy barrier. Consequently, the ion migration current density results in the sum of drift and diffusion components:

$$J_{ion} = J_{ion,drift} + J_{ion,diff} \tag{2.4}$$

where $J_{ion,drift}$ is simply the Mott-Gurney law (equation 2.3) and $J_{ion,diff}$ is:

$$J_{ion,diff} = zq\frac{dC}{dy}a^2\nu_0 \cdot \exp\left(-\frac{\Delta W_A}{k_B T}\right) \cdot 2\cosh\left(\frac{zq\mathcal{E}a}{2k_B T}\right) \tag{2.5}$$

Electron transport

Electronic conduction mechanisms in dielectric films [47, 51] can be classified as:

- Electrode-limited conduction mechanisms
    1. Schottky emission
    2. Fowler-Nordheim (FN) tunneling
    3. Direct tunneling
    4. Thermoionic-field emission
- Bulk-limited conduction mechanisms
    1. Ohmic conduction
    2. Space-Charge-Limited (SCL) conduction
    3. Trap-assisted conduction
        (a) Trap-to-trap tunneling, also called Trap-Assisted Tunneling (TAT)
        (b) Poole-Frenkel (PF) emission

Prior studies on the electron transport in filamentary TaO_x/HfO_x ReRAM device demonstrated that it is merely attributed to trap-assisted conduction mechanisms [39]. Consequently, for the purposes of this dissertation, the following discussion focuses only on TAT and PF emission, illustrated in Fig.2.12 as (a) and (b) respectively.

Trap-assisted transport processes are characteristic of substoichiometric, impurities-rich or amorphous dielectrics: in general, crystalline defects induce the presence of trap states within the band gap. Localized trap states can take part to the conduction by trapping and releasing electrons. The difference between TAT and PF emission rests on where electrons are released.

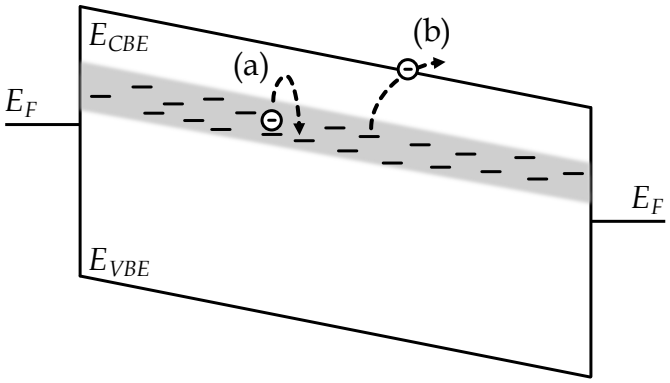


Figure 2.12: Schematic illustration of trap-assisted conduction mechanisms. (a) Trap-to-trap tunneling and (b) Poole-Frenkel emission. Redrawn from [39].

- Trap-to-trap (or trap-assisted) tunneling: electrons hop only among trap states, so they never jump into the conduction band. Moreover, TAT processes can occur in different regimes, depending on how far trap states are from each other. Low concentration of defects are associated to weak percolative conduction paths, while by increasing the defect concentration, the conduction trap centers density increases as well, falling in the weakly localized Variable Range Hopping (VRH) regime or more strongly localized Nearest-Neighbour Hopping (NNH) [47].
- Poole-Frenkel emission, also called internal Schottky emission, involves trapped electrons that are thermally excited and released in the conduction band. This effect typically occurs at high temperatures and high fields [51]. However, the temperature and the field required to activate the PF process depend also on the characteristic energy of localized defect states, translated into the energy difference between the Conduction Band Edge (E_{CBE}) and the energy of trap levels (ϕ_T).

Current densities (J_e) for TAT and PF processes can be described by the following equations:

$$J_e = q n_e a_e \nu_e \cdot \exp\left(-\frac{\Delta E_A}{k_B T}\right) \cdot 2\sinh\left(\frac{q\mathcal{E}a}{2k_B T}\right) \tag{2.6}$$

$$J_e = q\mu_e n_e \mathcal{E} \cdot \exp\left[\frac{\phi_T + \sqrt{q^3\mathcal{E}/\pi\epsilon_0\epsilon_r}}{k_B T}\right] \tag{2.7}$$

In equation 2.6 (TAT), n_e is the density of electronic states, $\mathcal{E}$ the electric field across the dielectric, a_e the average hopping distance, ν_e the frequency of thermal vibration of electrons in trap sites and E_A the average energy barrier between hopping sites.
In equation 2.7 (PF), μ_e is the electron mobility in conduction band, ϵ_0 the permittivity in vacuum and ϵ_r is the dielectric constant.

Temperature dynamics

Thermal transients throughout resistive switching processes are accurately modeled by the general Newton's cooling law [52, 53, 54], which describes the rate of heat loss from the warm object to the cooler surrounding. This general physics law states that the rate of heat loss is proportional to the temperature difference between the object and its surrounding and additional external contributions, such as the Joule heating, oppose to the cooling phenomena. Mathematically, the Newton's law of cooling reads:

$$C_{th} \cdot \frac{dT}{dt} = I_e \cdot V_A - \left(\frac{T - T_0}{R_{th}}\right) \tag{2.8}$$

where C_{th} is the thermal capacitace of the object, R_{th} the thermal resistance, I_e the electronic current flow causing Joule heating phenomena, V_A the applied voltage across the object, T the time-varying temperature and T_0 the reference temperature of the surrounding.

The thermal capacitance C_{th}, also known as heat capacity, quantifies the amount of heat required to increase the temperature by a certain amount, i.e. it represents the ability of the object to store thermal energy. C_{th} is associated merely to material properties and it is computed as:

$$C_{th} = c_p \cdot m$$

where c_p is the specific heat capacity and m the mass of the object.
The thermal resistance R_{th} measures the object ability to impede the temperature increase and it can be computed as:

$$R_{th} = \frac{l}{A} \cdot \frac{1}{\kappa}$$

where l is the length of the object, A the cross-sectional area of the object through which the heat flows and κ the thermal conductivity of the object material.
Thermal models described by the Newton's cooling law are equivalent to a simple RC circuit [52] in the thermal domain (Fig.2.13), whose time constant is:

$$\tau_{th} = C_{th} \cdot R_{th}$$

τ_{th} quantifies how fast is the heating up/cooling down transient.

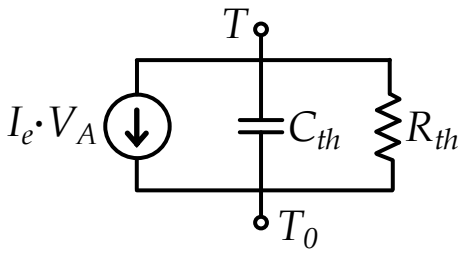


Figure 2.13: RC circuit in the thermal domain associated to the Newton's cooling law with Joule heating as external heat contribution.

2.3 Numerical approaches for physical modeling

Broadly speaking, physical modeling of ReRAM devices involves intricate systems of equations including differential and non-linear problems. It is also common to have dependent variables appearing in multiple equations. These complexities make impractical the analytical derivation of problem solutions. As a result, numerical approximations of such systems are essential to solve them with reasonable accuracy.
The core of numerical methods applied to physical problems is the discretization of governing equations, such as to translate them in manageable computational domains. Furthermore, iterative algorithms and numerical solvers offer a versatile way to compute the solution of discrete problems.

The intent of this section is to present 2 numerical methods to discretize and solve Ordinary Differential Equations (ODEs) and an iterative approach for non-linear systems of equations. The derivation of the discretization methods for ODEs can be found in [55].

2.3.1 Discretization of ordinary differential equations

The physics-based compact model derived in chapter 3 is a time-dependent problem related to multiple ODEs, i.e. differential equations containing a single independent variable. Initial values problem are treated in the mathematical model of this work, so the solution at time $t = 0$ must be known.
The general form of a 1D ODE is:

$$\dot{y} = \frac{\mathrm{d}f}{\mathrm{d}t} = f(x, t)$$

with initial condition $f(x, t = 0) = y_0$.
There are several approaches to discretized ODEs [55]. In the following, the two approaches used in this model are reported: the implicit Euler and the Crank-Nicolson (or trapezoidal rule) methods. They are chosen as discretization rules because of their unconditionally stable solution. In both cases it is necessary to discretize the time domain with finite steps Δt, where the general time t_i is:

$$t = \sum_{i=0}^{n} \Delta t_i$$

Euler method (implicit)

The implicit Euler method consist in approximating the time-derivative of the unknown as its finite difference within the discrete time interval.

$$\dot{y} = \frac{\mathrm{d}f}{\mathrm{d}t} = f(x, t) \approx \frac{f_i - f_{i-1}}{\Delta t_i}$$

therefore if the solution at the time t_i is known:

$$f_{i+1} = f_i + \Delta t_i \cdot f(x_{i+1}, t_{i+1})$$

and this formula can be applied progressively because the problem starts from known initial condition $f(x, t = 0)$ or $f_{i=0}$ in the discrete time domain.

Crank-Nicolson method

The Crank-Nicolson method, also known as trapezoidal rule is based on the approximation of the function f between 2 finite points of the discrete time domain as its trapezoidal evolution. This approach derives from the trapezoidal rule to compute integrals: its inverse rule is the Crank-Nicolson method. Of particular significance is the stability of the solutions found with this approach. Dealing with linear ODE problems, the trapezoidal rule leads to unconditionally stable solutions, which is essential for simulations. Only linear ODEs are involved in this physical modeling, i.e. the independent variables appear with the power of 1 in the differential problems.
Considering the following 1D problem,

$$\dot{y} = \frac{\mathrm{d}f}{\mathrm{d}t} = f(x, t)$$

the trapezoidal rule reads:

$$f_{i+1} = f_i + \frac{\Delta t_i}{2} \cdot (\dot{y}_i + \dot{y}_{i+1})$$

Also the Crank-Nicolson method can be applied progressively as the problem starts from known initial condition $f(x, t = 0)$ or $f_{i=0}$ in the discrete time domain, equally to the Euler approach.

2.3.2 Newton-Raphson numerical solver for non-linear systems

Non-linear systems arise when the governing equations exhibit non-linear relationship involving the independent variable. Numerical solvers for non-linear systems are often needed because many physical phenomena exhibit non-linear behavior: the complex interplay between different physical quantities and irregularities in the underlying physics make the problem impossible to be solved analytically.
Newton-Raphson method is particularly useful to solve such equation systems because it allows to find the solution with an iterative approach [56]. It consists in making an initial guess solution that is refined iteratively, until convergence within a specified tolerance is reached.
The Newton-Raphson method for a system of n equations ($f_{1,2,...,n}$) with n unknowns ($x_{1,2,...,n}$) is generalized in the following. The system must be written in the form $\bar{\mathcal{F}}(x) = 0$, where $\bar{\mathcal{F}} : \mathbb{R}^n \to \mathbb{R}^n$:

$$\bar{\mathcal{F}}(x) = 0 \to \begin{cases} f_1(x_1, x_{,2}, ..., x_n) = 0 \\ f_2(x_1, x_{,2}, ..., x_n) = 0 \\ \vdots \\ f_n(x_1, x_{,2}, ..., x_n) = 0 \end{cases}$$

Suppose to have an educated initial guess $\bar{\mathcal{X}}_0$, then $\bar{\mathcal{F}}$ is approximated by linearization around $\bar{\mathcal{X}}_0$:

$$\bar{\mathcal{F}}(x) \approx \bar{\mathcal{F}}(\bar{\mathcal{X}}_0) + \mathcal{J}(\bar{\mathcal{X}}_0) \times [(\bar{\mathcal{X}}) - \bar{\mathcal{X}}_0]$$

where $\bar{\mathcal{X}}$ is the vector containing the n solutions $x_{1,2,...,n}$ and $\mathcal{J}$ is the $n \times n$ Jacobian matrix whose entries are the combination of partial derivatives evaluated in $\bar{\mathcal{X}}_0$

$$\mathcal{J} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

therefore the first iteration to find the vector of the unknowns is:

$$\bar{\mathcal{X}} = \bar{\mathcal{X}}_0 - \mathcal{J}(\bar{\mathcal{X}}_0)^{-1} \times \bar{\mathcal{F}}(\bar{\mathcal{X}}_0)$$

with $\Delta\mathcal{X} = \bar{\mathcal{X}} - \bar{\mathcal{X}}_0$. By using $\bar{\mathcal{X}}$ as $\bar{\mathcal{X}}_0$ for the second iteration, the numerical solver can be repeated to find another $\bar{\mathcal{X}}$ which is closer to the real solution. The process can be iterated until $\Delta\mathcal{X}$ does not satify the numerical tolerance condition:

$$||\Delta\mathcal{X}|| < \epsilon_{tol}$$

where ϵ_{tol} is a small value chosen as numerical tolerance for the specific mathematical case.

Since it affects the method's effectiveness, it is crucial to carefully choose the initial guess. Moreover, the Jacobian matrix computation critically impact on the convergence towards the true solution, as it determines the updates of the unknowns vector. If poorly approximated, the entries of the Jacobian matrix might lead to slow convergences of the method, hence increasing the computational cost of the solver.

For time-dependent problems, the Newton-Raphson method can be employed to find numerically the solution at each time step. In this work, the ODEs discretization approaches are combined with a Newton-Raphson iterative solver to derive the time evolution of a physical problem.
All the simulations addressed in this work are implemented in MATLAB R2023b ©.

3 Modeling and results

This chapter is devoted to discuss the development of a physics-based compact model aimed to describe the resistive switching processes in ReRAM devices. The ReRAM technology studied in this dissertation is based on a stack of Conductive-Metal-Oxide (CMO) and HfO_x as active materials between electrodes, where the CMO is a Transition Metal Oxide (TMO) that meets specific constraints regarding its electro-thermal properties [39]. Specifically, in the first section of this chapter (3.1) it is explained how the theoretical fundamentals reported in 2.2 are applied to build a compact model based on well known electronic/thermal conduction mechanisms and ion kinetics in oxides. In section 3.2, the simulation results and the experimental data of electrical characterization are compared to evaluate the accuracy of the model.

3.1 Compact model

3.1.1 Physics of resistive switching dynamics

The approved interpretation about the resistive switching in Valence Change Memories (VCMs) consists in the migration of oxygen ions, also described as vacancies transport, causing a local valence alteration [57]. Similarly, the following model for analog filamentary CMO/HfO_x ReRAM devices assumes that a redistribution of defects in the CMO causes a modulation of the resistivity in that layer [39, 26]. In particular, the defects taken into account are oxygen vacancies, belonging to the class of lattice point defects.

Oxygen vacancies

A vacancy is a missing atom in a lattice site, as shown in Fig.3.1, hence, oxygen vacancies are missing oxygen ions in lattice sites that should be occupied by oxygens.
In Kröger-Vink notation [58], an oxygen vacancy is represented as $V_O^{\cdot\cdot}$, where the point defects species is V (vacancy in this case), the subscript is the original lattice site considering the perfect crystal (O stands for oxygen) and the superscript is the negative ($^{\cdot}$), positive ($^{\prime}$) or null (^{x}) net charge. Since oxygen ions are negatively charged with 2 unpaired electrons, $V_O^{\cdot\cdot}$ have double positive net charge with respect to the lattice.
Vacancies are not existing species, so a rigorous description should not consider them as migrating entities because there is no physical matter to move: nevertheless, the description of vacancy migration (complementary movement of self-interstitial atoms in the opposite lattice site direction) is accepted as long as there is no atom exchange at the interface between different materials.
Other types of point defects are extrinsic interstitial atoms, substitutional atoms and Frenkel pairs (see Fig.3.1). All of them are not considered: in this model, the active materials responsible of resistive switching are not supposed to have atoms that do not belong to their crystalline structure (extrinsic defects), while Frenkel pairs are excluded due to recombination effects.

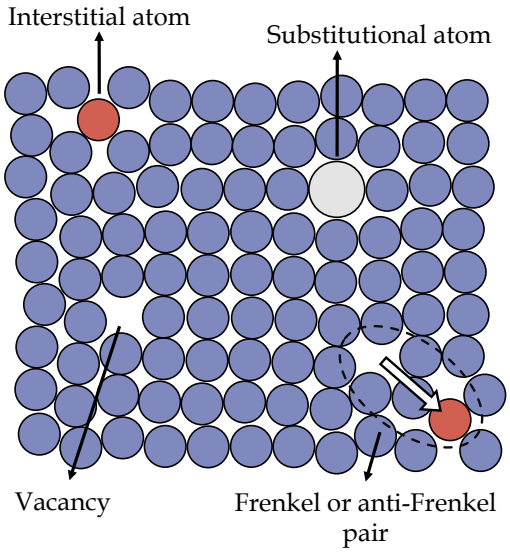


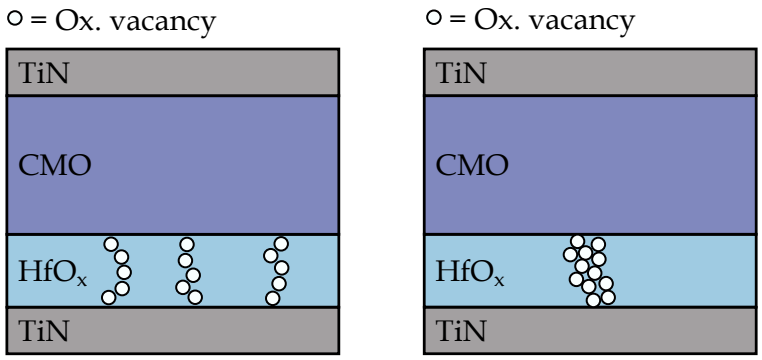
Figure 3.1: Classification of point defects in crystalline structures.

The distinction between Frenkel pairs and anti-Frenkel paiars relies on the charge of the vacancy left by the migration of an ion to the adjacent interstitial lattice site, so oxygen vacancies-oxygen interstitial ($V_O^{\cdot\cdot}$-O_i'') pairs are anti-Frenkel defects. The latter can exist in oxides, but they are not taken into account in this work because studies on anti-Frenkel $V_O^{\cdot\cdot}$-O_i'' pairs in TMOs [59] reveled that they are characterized by very fast recombination times ($< 1\ ps$), i.e. they are not stable. To summarize, the migration of ionic oxygen species is depicted in terms of $V_O^{\cdot\cdot}$ migration and this allows to discriminate them from anti-Frenkel pairs.

Interpretation of resistive states

As explained in sections 1.1-1.3, ReRAMs operating principle is based on the resistance change caused by the alteration of the conductive properties of an active material between two electrodes. The active materials in the ReRAM device of this study are TaO_x and HfO_x, stacked as illustrated in appendix A. The substoichiometric TaO_x is chosen as representative CMO layer. However, the model proposed in the next section holds for all the CMOs that fulfill the electro-thermal requirements listed in section 3.1.2. Therefore, the general interpretation of resistive states is explained in terms of CMO.
HfO_x conductive properties change only during the electro-forming step, where the substoichiometric HfO_x is further reduced (locally) and a conductive filament made of $V_O^{\cdot\cdot}$ is formed, bridging the CMO layer with the bottom electrode. A realistic interpretation of electro-formed oxygen deficient filaments consists in considering them as formed in multiple sites (Fig.3.2a), as reported in [60]. Nonetheless, for the purpose of this simulation model, it is sufficient to approximate the conductive filament as unique (Fig.3.2b), according to recommended simulation methods for resistive switching devices [35].
The filament is ohmic-like due to the high oxygen deficiency: the transport in the filament is dominated by the metallic phase with highly conductive percolation paths [61].

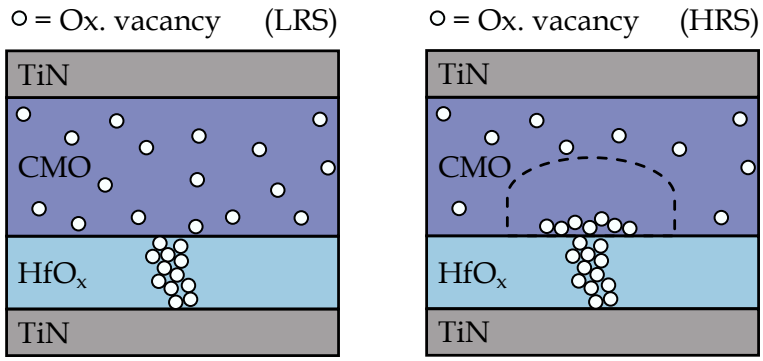
Despite the compact model proposed in this study does not take into account any spatial variable, it is crucial to establish the hypothesis regarding the arrangement of $V_O^{\cdot\cdot}$ in the device, because simulations results shall be construed accordingly.
The model takes for granted that the redistribution of $V_O^{\cdot\cdot}$ responsible for resistance change



(a) (b)

Figure 3.2: (a) Realistic interpretation of multiple oxygen vacancies filaments formation as a consequence of the electroforming step. (b) Unique filament approximation.

during SET and RESET processes, occurs in a sub-portion of the CMO layer: in [39, 40], it is shown that the electrostatic potential drops mainly in a dome-shaped region of the CMO on top of the conductive filament, therefore also the CMO sub-volume undergoing the change of oxygen vacancies concentration $N_{V_O^{\cdot\cdot}}$ is assumed to be a dome above the conductive filament. When the device is in LRS, $V_O^{\cdot\cdot}$ are homogeneously distributed in the CMO layer (see Fig.3.3a). They come from the partial reduction induced by the first step of the electroforming procedure. By applying a positive bias (V_A) to the top electrode, sufficiently high to generate a $V_O^{\cdot\cdot}$ migration, they drift towards the interface of the HfO_x and the dome is partially depleted of defects (RESET). Once the $N_{V_O^{\cdot\cdot}}$ in the CMO dome decreases (the dome is oxidized), the device is in HRS (see Fig.3.3b). From HRS to LRS (SET), a sufficiently strong negative bias applied to the top electrode allows to relocate $V_O^{\cdot\cdot}$ in the LRS configuration, thereby repopulating the dome of defects.



(a) (b)

Figure 3.3: Oxygen vacancies spatial arrangement interpretation in CMO/HfO_x ReRAM in LRS (a) and HRS (b).

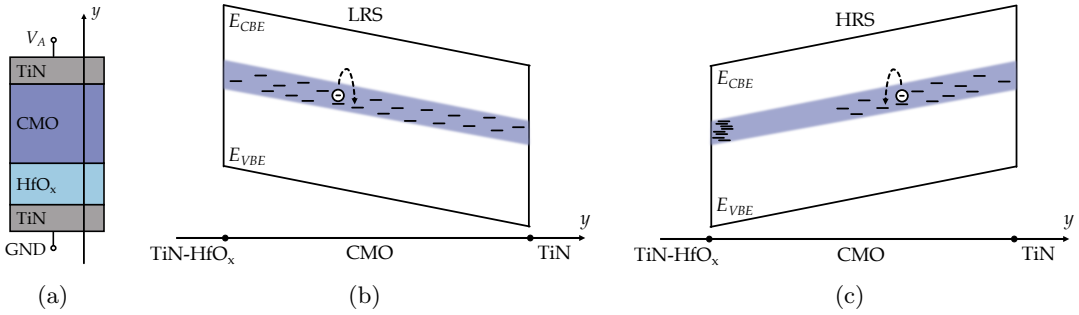
It is widely approved that the presence of $V_O^{\cdot\cdot}$ gives rise to defect trap states in the band gap of the oxide, changing its conductivity: they are treated as donor-like trap states [62]. In subsection 2.2.1, two types of electron conduction mechanisms are supposed, both assisted by the presence of defect trap states in the band gap of the oxides:

- Poole-Frenkel (PF) conduction, consisting in thermally excited electrons that are re-

leased from trap states to the conduction band, with the following drift towards the electrode

- Trap-Assisted Tunneling (TAT), where electrons are released from the Fermi level to a trap state , then tunneling through localized defect states towards the direction of the bias

The type of conduction mechanisms strongly depends on the energy distance between the defect traps state and the conduction band edge. For instance, TAT electron transport is characteristic of oxides with deep trap states. In this model, the PF process is discarded, as the redistribution of $V_O^{\cdot\cdot}$ occurs in the CMO, which is supposed to have middle gap defect states. This evaluation applies for many TMOs characterized by oxygen vacancies-related deep trap states [63]. Specifically, considering the case study of this work, it has been demonstrated that oxygen vacancies trap states are $\sim 2\ eV$ far from the valence band edge in TaO_x [64], i.e. they are spectrally located in middle of the band gap.
Following these considerations, Fig.3.4 depicts how the conduction through localized defect trap states (consequently, the resistance of the CMO) is affected by the position of such defects.



(a) (b) (c)

Figure 3.4: (a) Direction of the electronic conduction/ion migration. (b),(c) Sketches of spectral and spatial location of defects trap states in the band diagram for LRS and HRS respectively, redrawn from [39].

Since the LRS is associated to homogeneously distributed $V_O^{\cdot\cdot}$ in the whole CMO, regular TAT electron paths between the HfO_x conductive filament and top electrode exist, passing through $V_O^{\cdot\cdot}$ trap states (see Fig.3.4b). In the HRS the conduction bottleneck is the dome region, which is partially depleted of $V_O^{\cdot\cdot}$ defect states, hence TAT in such portion of CMO is limited by traps deficit (see Fig.3.4c).
As shown both in Fig.3.3b and Fig.3.4c, in the HRS the $V_O^{\cdot\cdot}$ are deemed to be accumulated at the interface without crossing the interface with the HfO_x: this hypothesis might not be valid anymore when strong RESET biases are applied to the device, which allow to overcome the potential barrier of oxygen interlayer exchange. The experimental data of SET/RESET quasi-static operation reported in 2.1 involve maximum biases of $\sim 1\ V$. Even if there are no data about the energy barriers for oxygen exchange at HfO_x/TaO_x interface, it is reasonable to assume that those biases are not sufficiently high to induce the interlayer migration.

In summary, this compact model relies on the following assumptions:

- the conductive filament is unique, ohmic-like and it does not alter during switching processes
- the point defects generated in the oxides by the electro-forming are only oxygen vacancies

- the electro-forming step is not taken into account
- oxygen vacancies introduce donor-like defects in the middle of the gap, far enough from conduction band edge to avoid thermally excited electrons for band transport
- the transport is based purely on Trap-Assisted-Tunneling
- the resistive switching is the consequence of oxygen vacancies redistribution in a dome-shaped sub-region of the CMO above the conductive filament

3.1.2 Equivalent circuit

The compact model presented in this section is based on the physical mechanisms illustrated in 2.2. The equations are coupled with an equivalent electrical circuit whose components are associated to different materials of the device stack. The model addressed in this dissertation holds for analog filamentary ReRAM devices based on Metal-CMO-HfO_x-Metal stack for a generic CMO that satisfies the assumption listed in subsection 3.1.1. Furthermore, the CMO must fulfill constrains regarding its electro-thermal properties in order to have CMO-HfO_x-based ReRAM devices with the same electrical properties. In particular, the electrical conductivity of the CMO (σ_{CMO}) and the thermal conductivity (κ_{CMO}) must be roughly one order of magnitude smaller than σ_{cf} and κ_{cf} [39].
A substoichiometric TaO_x is chosen as representative CMO layer.

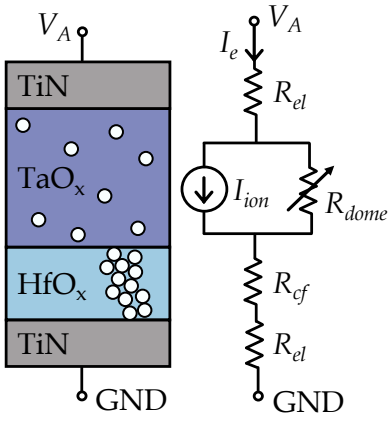


Figure 3.5: Equivalent circuit for the electrical model of CMO/HfO_x ReRAM with TaO_x as CMO.

The equivalent circuit model (Fig.3.5) refers to the fabricated TiN-TaO_x-HfO_x-TiN ReRAM device (see appendix A) after the electro-forming step, when an ohmic-like conductive filament is already present in the HfO_x layer. Moreover, the filament does not contribute to resistive switching processes, i.e. its resistance is invariant: FEM simulations of the same device [39, 40] showed that heat and electric field confinements occur in the TaO_x dome on top of the filament, so in the HfO_x there is no electro-thermal driving force to displace $V_O^{\cdot\cdot}$ belonging to the oxygen deficient conductive filament. Accordingly, the filament is modeled as a series resistor R_{cf}, as well as the electrodes $2R_{el}$. The electrodes and the conductive filament resistances are computed through:

$$R = \frac{l}{A} \cdot \sigma^{-1} \tag{3.1}$$

Regarding the electrodes, the geometrical parameters l_{el} and A_{el} are respectively the thickness of the TiN and the area of the cell, while the TiN electrical conductivity σ_{TiN} is

the same used in [39].
The conductive filament is approximated as cylindrical, so l_{cf} coincides with the HfO_x thickness and the area is computed as $A_{cf} = \pi \cdot r_{cf}^2$, where r_{cf} is the radius of the cylindrical conductive filament. r_{cf} and σ_{cf} are taken from [39].
R_{dome} is the resistance representing the TaO_x layer. The subscript "*dome*" specifies that the resistance variation is computed taking into account the hypothesis of the model explained in section 3.1.1: only the dome region undergoes the $V_O^{\cdot\cdot}$ redistribution, varying its resistance. The concentration of oxygen vacancies in the dome ($N_{V_O^{\cdot\cdot}}$), i.e. the number of $V_O^{\cdot\cdot}$ per unit volume, is used as state variable: its variation causes a change of conduction properties. The following Ordinary Differential Equation (ODE) is used to compute the concentration variation during time:

$$\frac{dN_{V_O^{\cdot\cdot}}}{dt} = -\left(\frac{1}{qzV_{dome}}\right) \cdot I_{ion} \tag{3.2}$$

q is the elementary charge, z is the defect charge number and V_{dome} is the volume of the dome [39], used to normalize the rate of defects displacement. Since the $V_O^{\cdot\cdot}$ are defects associated to the displacement of a doubly-charged oxygen ion, the charge number is $z = +2$. The ionic current I_{ion} determines the change of $N_{V_O^{\cdot\cdot}}$ and it is described as:

$$I_{ion} = (J_{ion,drift} - J_{ion,diff}) \cdot A_{dome} \tag{3.3}$$

where the drift and diffusion components are computed according to the hopping model derived in [50] for the homogeneous oxygen vacancies migration as a function of temperature and electric field:

$$J_{ion,drift} = zqN_{V_O^{\cdot\cdot}} a\nu_0 \cdot \exp\left(-\frac{\Delta W_A}{k_B T}\right) \cdot 2\sinh\left(\frac{zq\mathcal{E}a}{2k_B T}\right) \tag{3.4}$$

$$J_{ion,diff} = zq\frac{dN_{V_O^{\cdot\cdot}}}{dy} a^2\nu_0 \cdot \exp\left(-\frac{\Delta W_A}{k_B T}\right) \cdot 2\cosh\left(\frac{zq\mathcal{E}a}{2k_B T}\right) \tag{3.5}$$

The gradient induced by $V_O^{\cdot\cdot}$ migration, responsible of a non-null diffusion component, is approximated as the maximum value it could assume:

$$\frac{dN_{V_O^{\cdot\cdot}}}{dy} \approx \frac{N_{V_O^{\cdot\cdot}}^{LRS} - N_{V_O^{\cdot\cdot}}^{HRS}}{l_{CMO}/2} \tag{3.6}$$

The parameter A_{dome} in equation 3.3 is the cross-sectional area of V_{dome}, perpendicular to the conduction/migration direction (y in Fig.3.4a): this volume is assumed to be a semi-sphere with base area slightly larger (20 %) than A_{cf} ($A_{dome} = 1.44 \cdot A_{cf}$), so its cross-section is not unique. However, to keep the model as simple as possible and physically plausible at the same time, the realistic geometry of the dome is not considered. V_{dome} in rectangular shape approximation with $A_{dome} = 1.44 \cdot A_{cf}$ would have a thickness of 10.6 nm, roughly half of the TaO_x layer thickness.
In the equations 3.4 and 3.5, a is the hopping distance between adjacent oxygen sites, ν_0 the jump attempt frequency, ΔW_A the zero-field activation energy barrier for $V_O^{\cdot\cdot}$ migration, k_B the Boltzmann constant, $\mathcal{E}$ the electric field and T the average temperature in the TaO_x.
The electric field generated by the positive (RESET) or negative (SET) applied bias to the top electrode is expressed as:

$$\mathcal{E} = \frac{V_A}{l_{CMO}} \tag{3.7}$$

Equation 3.5 is based on considerations about the electrical conductivities of the involved materials [39]:

$$\sigma_{\mathrm{TiN}} = 5 \cdot 10^5 \ Sm^{-1} > \sigma_{cf} = 4.2 \cdot 10^4 \ Sm^{-1} > \sigma_{\mathrm{TaO_x}} \sim 2 \cdot 10^3 \ Sm^{-1}$$

Accordingly, most of the applied electric potential is confined in the least conductive material. i.e. the $\mathrm{TaO_x}$ ($l_{CMO} = l_{\mathrm{TaO_x}}$ here). This holds in general for all the CMOs that satisfy the electro-thermal constraints in CMO/$\mathrm{HfO_x}$ ReRAM [39, 40].

In order to describe the current flow in the device (I_e), the TAT process through midgap defects states is modeled as a Mott-Gurney law [48] for electron hopping.

$$I_e = A_{dome} q N_e a_e \nu_e \cdot \exp\left(-\frac{\Delta E_A}{k_B T}\right) \cdot 2\sinh\left(\frac{q\mathcal{E}a_e}{2k_B T}\right) \tag{3.8}$$

Here, a_e is the average trap-to-trap distance in the defects-rich $\mathrm{TaO_x}$, ν_e the electron attempt frequency referring to the hopping between trap states, ΔE_A the zero-field hopping energy barrier and N_e the density of electronic states. Considering that $\mathrm{V_O^{\cdot\cdot}}$ are treated as donor-like defects with $z = +2$, each of them induces the presence of 2 trap states spectrally located in the midgap. Some trap states can be occupied, so not all of them take part to the conduction and N_e can be approximated as:

$$N_e \approx \beta \cdot z \cdot N_{\mathrm{V_O^{\cdot\cdot}}} \tag{3.9}$$

where β is an arbitrary scaling factor ($0 < \beta < 1$) to take into account the unavailable trap states for the conduction.
It is clear that the conduction mechanism represented by equation 3.8 is non-linear: the relation between the electronic current I_e and the electric field $\mathcal{E}$ exhibits a sinh proportionality, as the $\mathcal{E}$ lowers/increases the hopping energy barrier by $\mp q\mathcal{E}a_e/2$ for forward/backward trap-to-trap tunneling. Therefore the non-linear variation of R_{dome} caused by the change of $N_{\mathrm{V_O^{\cdot\cdot}}}$ (so N_e too) is computed as:

$$R_{dome} = \frac{V_A}{I_e} - R_{cf} - 2Rel = \frac{V_A}{I_e} - R_{series} \tag{3.10}$$

Finally, the Newton's cooling law is used to compute the time evolution of the average temperature (T) in the $\mathrm{TaO_x}$ dome:

$$C_{th} \cdot \frac{dT}{dt} = I_e \cdot V_A - \left(\frac{T - T_0}{R_{th}}\right) \tag{3.11}$$

C_{th} and R_{th} are the thermal capacitance and thermal resistance of dome respectively, whereas T_0 is the room temperature taken as a reference when the device is not heated up as a consequenc of Joule heating. Both C_{th} and R_{th} are computed taking into account thermal and electrical properties of the dome. First of all, it is essential to estimate the O/Ta fraction depending on $\sigma_{\mathrm{TaO_x}}$. In [65], it has been reported the electrical conductivity measurement of $\mathrm{TaO_x}$ with $0 < \mathrm{x} < 2.36$, while according to this study O/Ta ≈ 1.8 when $\sigma_{\mathrm{TaO_x}} \sim 2 \cdot 10^3 \ Sm^{-1}$. In addition, the computed stoichiometry is feasible considering the Grazing Incidence X-Ray Diffraction (GIXRD) data [26] about the same fabricated device modeled in this work. The extracted oxygen percentage is used to compute the amount of $\mathrm{TaO_x}$ in the dome expressed in moles through the Molar Mass (MM) of its chemicals:

$$MM = MM_{\mathrm{TaO_x}} + \mathrm{x} \cdot MM_O = (180.94788 + 1.8 \cdot 15.999)\, g/mol = 209.75 \ g/mol$$

therefore the moles of $\mathrm{TaO_x}$ in V_{dome} are:

$$M_{\mathrm{TaO_x}} = \frac{\rho_{\mathrm{TaO_x}} \cdot V_{dome}}{MM} = \frac{9.3\ g/cm^3 \cdot 3 \cdot 10^{-23}\ m^3}{209.75\ g/mol} = 1.33 \cdot 10^{-18}\ mol$$

where $\rho_{\mathrm{TaO_x}}$ is taken from the reference database [66]. The specific heat (c_p) of $\mathrm{TaO_x}$ as a function of temperature was measured in [67, 68]. To reduce the computational cost of the simulation, a unique value of c_p is chosen and it is computed as the average in the temperature range in which this device can work. Consequently, $c_p \sim 160\ J/(mol\ K)$ in $300\ K < T < 2000\ K$, so the thermal capacitance is computed as:

$$C_{th} = c_p \cdot M_{\mathrm{TaO_x}} = 2.13 \cdot 10^{-16}\ J/K$$

The thermal resistance could be computed from the thermal conductivity (κ) of the material subjected to heating:

$$R_{th} = \frac{l}{A} \cdot \frac{1}{\kappa_{\mathrm{TaO_x}}}$$

However, the previous relation for R_{th} holds only when A_{dome} is constant along the accounted thickness and this is not valid for a semi-spherical dome. R_{th} is a delicate parameter in the thermal model represented by equation 3.11 and any approximation might lead to wrong results, so it is treated as a fitting parameter to match the quasi-static I-V sweep experimental data reported in section 2.1. The extracted parameter of R_{th} is physically plausible for the following reasons.

- $R_{th}^{'}$ computed considering a (poor) approximation for the dome geometry as heated volume:
$$R_{th}^{'} = \frac{l_{dome}}{A_{dome}} \cdot \frac{1}{\kappa_{\mathrm{TaO_x}}} = 3.75 \cdot 10^6\ K/W$$

- $R_{th}^{''}$ computed for a regular cross-section considering the whole layer of $\mathrm{TaO_x}$ as heated volume:
$$R_{th}^{''} = \frac{l_{\mathrm{TaO_x}}}{A_{el}} \cdot \frac{1}{\kappa_{\mathrm{TaO_x}}} = 4.25 \cdot 10^5\ K/W$$

where $\kappa_{\mathrm{TaO_x}} \sim 1\ W/(m\ k)$ [69]. $R_{th}^{'}$ and $R_{th}^{''}$ are two extreme cases and the value found to match the experimental data falls within this range:

$$R_{th}^{'} < R_{th} = 6.3795 \cdot 10^5\ K/W < R_{th}^{''}$$

Symbol	Value	Symbol	Value	Symbol	Value
l_{el}	$20\ nm$	A_{el}	$(200\ nm)^2$	σ_{TiN}	$5 \cdot 10^5\ Sm^{-1}$
l_{cf}	$3.5\ nm$	r_{cf}	$25\ nm$	σ_{cf}	$4.2 \cdot 10^4\ Sm^{-1}$
$l_{\mathrm{TaO_x}}$	$17\ nm$	$\kappa_{\mathrm{TaO_x}}$	$1\ W/(mk)$	$\sigma_{\mathrm{TaO_x}}$	$2 \cdot 10^3\ Sm^{-1}$
κ_{cf}	$23\ W/(mK)$	V_{dome}	$3 \cdot 10^{-23}$	A_{dome}	$1.44 \cdot \pi r_{cf}^2$
R_{series}	$R_{cf} + 2R_{el}$	z	$+2$	β	0.5
a	$0.4\ nm$	a_e^{LRS}	$0.75\ nm$	a_e^{HRS}	$0.88\ nm$
ν_0	$4 \cdot 10^{12}\ Hz$	ν_e	$2 \cdot 10^{13}\ Hz$	ΔE_A^{LRS}	$65\ meV$
ΔE_A^{HRS}	$82\ meV$	ΔW_A^{RESET}	$1.45\ eV$	$\Delta W_{A,0}^{SET*}$	$0.84\ eV$
T_0	$293\ K$	C_{th}	$2.13 \cdot 10^{-16}\ J/K$	R_{th}	$6.38\ K/W$

Table 3.1: Simulation parameters.

$\Delta W_{A,0}^{SET}*$ (see table 3.1) has the pedix "*0*" because is assumed to increase linearly with the state variable $N_{V_O^{\cdot\cdot}}$ within the range $[\Delta W_{A,0}^{SET};\ \Delta W_A^{RESET}]$. The reason of the assumption will be explained in the next section.
All the simulation parameters listed in Table 3.1 are physically reasonable. Geometrical parameters refer to the structure of the fabricated TiN-TaO_x-HfO_x-TiN ReRAM device (see appendix A).

3.1.3 Algorithm implementation

Equations from 3.1 to 3.10 are all interlinked between each other, with vacancy concentration, temperature, electric field and currents that appear in multiple expressions. Moreover the system of equations is non-linear, including ODEs and variables as argument of functions. For this reason the system is solved numerically by applying the iterative Newton-Raphson method explained in section 2.3. Nevertheless, not all the equations must be included inside the numerical solver, i.e. their solution can be found before or after it. The minimum size of the numerically solved system is 4, including the equations 3.2, 3.3, 3.8 and 3.11. It returns the solution $[N_{V_O^{\cdot\cdot}},\ I_{ion},\ I_e,\ T]$ for each applied bias V_A. Equation 3.1 is used only to compute simulation parameters, such as $R_{cf},\ R_{el}$, while equations 3.4, 3.5, 3.6, 3.9 are implicitly included in the system as part of other equations. The electric field (equation 3.7) and the resistance R_{dome} (equation 3.10) can be computed before and after the numerical solver respectively.
The ODEs 3.2 and 3.11 are discretized in time such as to have them in a "numerical-friendly" form, according to the Crank-Nicolson rule for equation 3.2 and the Euler rule for equation 3.11:

$$\frac{dN_{V_O^{\cdot\cdot}}}{dt} = -\left(\frac{1}{qzV_{dome}}\right)\cdot I_{ion} \quad \rightarrow \quad N_{V_O^{\cdot\cdot}}^{i} = N_{V_O^{\cdot\cdot}}^{i-1} - \Delta t\cdot\left(\frac{1}{qzV_{dome}}\right)\cdot\frac{\left(I_{ion}^{i}+I_{ion}^{i-1}\right)}{2}$$

$$C_{th}\cdot\frac{dT}{dt} = I_e\cdot V_A - \left(\frac{T-T_0}{R_{th}}\right) \quad \rightarrow \quad T^i = T^{i-1} + \frac{\Delta t}{C_{th}}\cdot\left[I_e\cdot V_A - \left(\frac{T^i-T_0}{R_{th}}\right)\right]$$

where i is the index used to discriminate different biases.

To initiate the solver, the solution $[N_{V_O^{\cdot\cdot}},\ I_{ion},\ I_e,\ T]$ needs an educated guess regarding the starting resistance state, when no bias is applied: without bias there is no electric field, so I_{ion} and I_e are null and $T = T_0$, while $N_{V_O^{\cdot\cdot}}$ has a non-null value corresponding to the static HRS or LRS (depending if the simulation starts with a SET or a RESET). Considering the $N_{V_O^{\cdot\cdot}}$ in static regime computed by Falcone et al. [39], a well-conditioned guess is $N_{V_O^{\cdot\cdot}} \sim 10^{26}\ m^{-3}$.
Then the solver can start the iterative computation updating the solution. The Jacobian with all the combinations of derivatives is computed analytically, so the solver can run to find the solution of the non-linear system for each applied bias V_A^i. The detailed matrix form of the non-linear system and its Jacobian is given in appendix C.
For each i-th bias the iterative Newton-Raphson method is applied until the convergence condition based on an pre-established numerical tollerance is satisfied. If convergence is not reached the solver decreases the discrete time step (Δt) and the non-converging solution is discarded. The iterative procedure is stopped when a maximum number of iterations without convergence is reached. After convergence, the solution is stored and a new bias (V_A^{i+1}) simulation is run.

Depending on the type of $V(t)$ characteristic to be simulated, the condition to conclude the algorithm changes.

- For quasi-static I-V sweep simulation, the algorithm is "voltage-controlled", since it runs for SET and RESET and it is stopped when $V_A^i = V_{Stop}$ (see Fig.3.6). The algorithm corresponds to the simulation of a triangular voltage sweep with a constant Sweep Rate (SR).

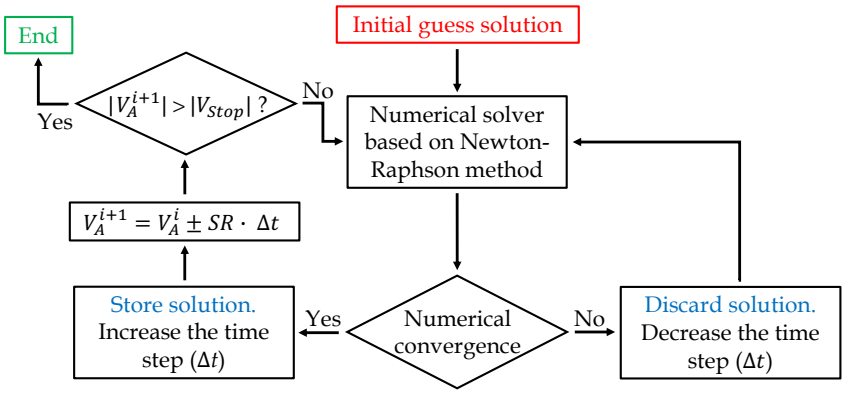


Figure 3.6: Flowchart for quasi-static I-V sweep simulation.

- For pulse response simulation, the algorithm is "time-controlled", as it is stopped when the defined duration of the pulse sequence is reached (see Fig.3.7).

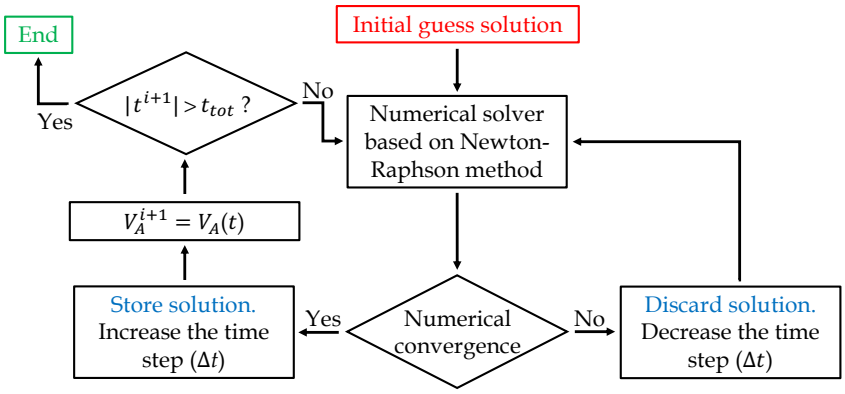


Figure 3.7: Flowchart for pulse response simulation.

The algorithm explained in this section is taken from [70].

3.2 Compact model validation

In this section, the simulation results are discussed and linked to the physics of resistive switching mechanisms. To check that the model is realistic and accurate, its simulation output is also compared with the experimental data about electrical characterizations both in the quasi-static and AC domain.

3.2.1 Quasi-static voltage sweep model

According to recent studies on physical compact modeling of VCM devices [70], the diffusion component $J_{ion,diff}$ appearing in equation 3.5 allows to model the finite retention of the

device under null bias condition. Moreover, Noman et al. [50] showed that a drift-only model can not explain the retention in memristive devices involving ionic motion. However, in this work the retention/endurance simulations are not addressed, so to reduce the computational cost of the simulation, the drift-diffusion equation 3.3 is simplified as a drift-only one: the Noman et al. model [50] reduces to the Mott-Gurney law for ion hopping [48]:

$$I_{ion} \approx J_{ion,drift} \cdot A_{dome}$$

Binary state modeling

The algorithm depicted in Fig.3.6 is applied to emulate the triangular SET/RESET sequence used to obtain the quasi-static I-V sweep experimental data. The sweep parameters used in the model are the same as in the experiment, i.e. $V_{Stop}^{SET} = -0.9\ V$, $V_{Stop}^{RESET} = +1.1\ V$ and $\mathrm{SR} = 0.1\ Vs^{-1}$, resulting in a sweep lasting 40 s as shown in Fig.3.8.

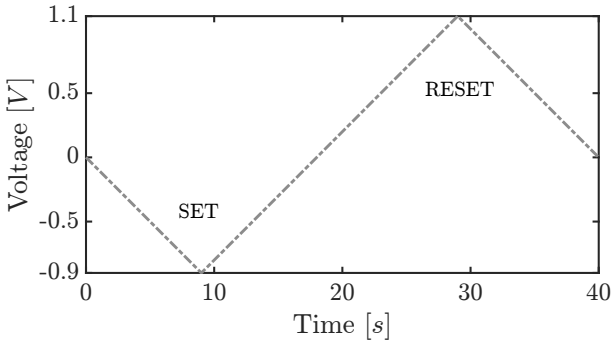


Figure 3.8: Voltage-time characteristic employed to simulate the I-V sweep.

The sign of the electrostatic voltage of the triangular sweep refers to the top electrode of the device.
For a first validation of the model, the simulated I-V characteristic is superimposed to 10 cycles of experimental sweep data: as shown in Fig.3.9a and 3.9b (reporting the same plot in different scales), the data overlap with high accuracy.

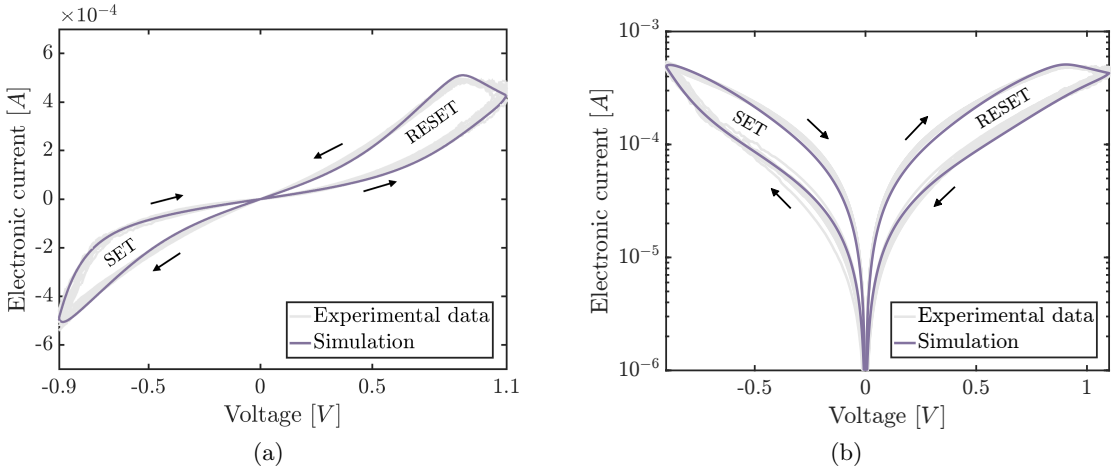


Figure 3.9: Measured and simulated clockwise I-V characteristics of TaO_x/HfO_x-based ReRAM device in linear (a) and logarithmic (b) scales.

The resulting I-V plots confirm that the hopping transport is appropriate to describe the

conduction in TaO_x/HfO_x-based ReRAM device both during the static phase and during the resistive switching processes. This is in agreement with the study on the TaO_x stoichiometry linked to the dominating conduction mechanism in the defective material: Heisig et al. [71] demonstrated that the resistive switching in TaO_x can be attributed to a modulation of the stoichiometry (x) and hopping transport dominates when the O/Ta ratio (x) is the range $\{0.75; 1.9\}$. This aligns with the computed stoichiometry in subsection 3.1.2. In this regime the resistivity is reported to be exponentially increasing with x. As illustrated in Fig3.10a, the resistance is not varying exponentially during the SET/RESET, suggesting that the modulation of stoichiometry in this device is only partial.
Fig.3.10b (blue curve) depicts the time evolution of $N_{V_O^{\cdot\cdot}}$, whose variation is assumed to occur in V_{dome}. Also the $N_{V_O^{\cdot\cdot}}(t)$ plot confirms that the modulation is partial because the ratio $N_{V_O^{\cdot\cdot}}^{LRS}/N_{V_O^{\cdot\cdot}}^{HRS}$ is less than 2, as demonstrated by Falcone et al. [39]. The gradual concentration change during the SET starts at $\sim 7\ s$, which coincides with $V_A \sim -0.7\ V$ according to the sweep rate used. During the RESET phase, when the dome is depleted of defects, the concentration change starts at $\sim 26\ s$, i.e. at $V_A \sim +0.8V$. Thus, the $V_O^{\cdot\cdot}$ redistribution occurs only during the resistive switching, even if the resistance is not constant outside the transition phases. The reason is that the hopping conduction modeled with equation 3.8 is non-linear by definition.

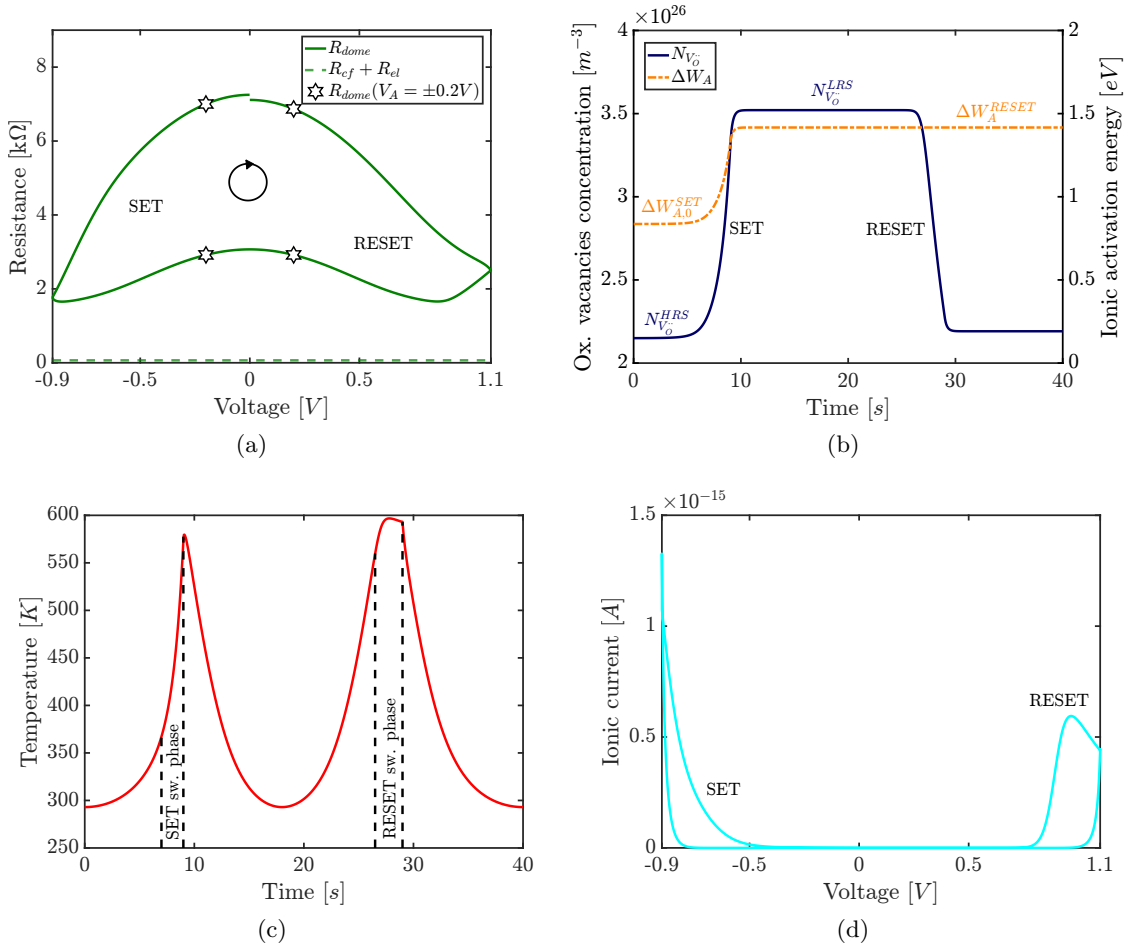


Figure 3.10: (a) *R-V* characteristic. (b) Time evolution of oxygen vacancy concentration and ionic migration energy barrier. (c) Time evolution of the average temperature in the dome. (d) Ionic current curing the SET/RESET sweep.

In [39] the vacancies migration energy barriers to initiate the switching differ between SET and RESET, complying with the following relation:

$$\Delta W_A^{SET} < \Delta W_A^{RESET}$$

Hence in this model, ΔW_A varies over the time span of the simulation, as shown in Fig.3.10b (orange curve). Based on the interpretation of $V_O^{\cdot\cdot}$ distribution (see Fig.3.3) in LRS/HRS, at the SET onset ($t \sim 7\ s$) the defects concentration gradient between the TaO_x/HfO_x interface and the partially depleted dome lowers the migration activation energy. While defects are relocated in the dome (restoring the LRS) ΔW_A increases up to ΔW_A^{RESET} because the migration is not thermodynamically favorable anymore when the gradient decreases. For this reason ΔW_A is assumed to increase linearly with $N_{V_O^{\cdot\cdot}}$ during SET and stays constant during the RESET. This assumption aligns with the study of Woo et al. [72], where the ion migration responsible of resistive switching is affected by the environment, i.e. by the oxygen/metal concentration ratio.
In Fig.3.10c, showing the average temperature in the dome over the sweep time, the opposite $T(t)$ trend is highlighted for SET/RESET switching phases. This is explained in terms of Joule heating feedback.

- Throughout the SET phase, the resistance decreases from HRS to LRS, leading to an increase of the electronic current flowing in the device. Due to Joule heating, the temperature increases exponentially because of the positive feedback between T and I_e: the SET is characterized by thermal runaway phenomena [29].
- During the RESET phase, T and I_e are in negative feedback because the increase of the resistance (from LRS to HRS) counteracts the current flow and the Joule heating. For this reason the temperature stays constant in the switching phase.

The simulated temperatures at the SET/RESET onset are respectively $T = 370\ K$ and $T = 560\ K$. These values are consistent with the temperature extracted in the prior study on electro-thermal FEM simulations for the same device [39], corroborating the considerations on the thermal model used in this work.
The drift ionic current sweep is plotted in Fig.3.10d, whereas Fig.3.11a/b show the drift-diffusion components during the SET/RESET switching time intervals. The diffusion ionic current is computed at the end of the simulation to check that it is negligible with respect to the drift one.

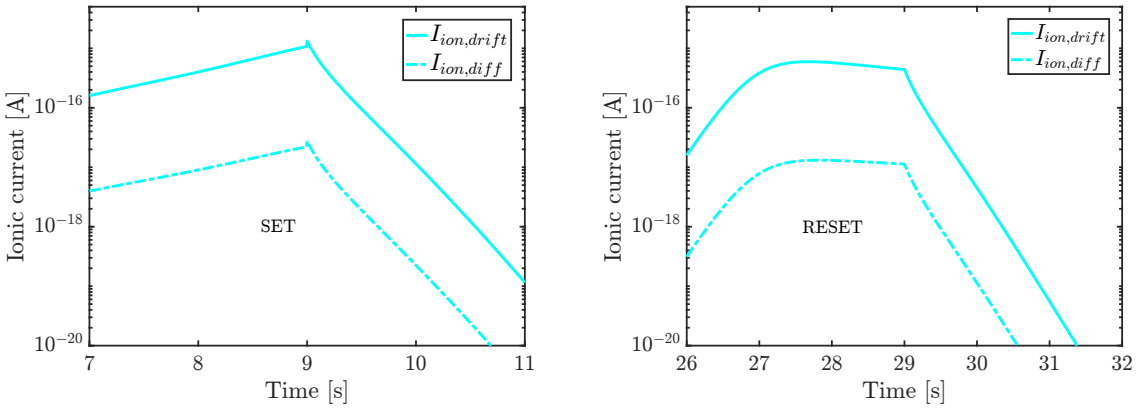


Figure 3.11: Ionic current drift-diffusion components during SET (a) and RESET (b) switching phases.

Despite $I_{ion,diff}$ is overestimated with equation 3.6 (the gradient is set at its maximum value), the drift component is 2 orders of magnitude larger. Moreover the modulation of $N_{V_O^{\cdot\cdot}}$ is limited in this device. Therefore $I_{ion,diff}$ can be neglected in single cycle I-V sweep simulations.

Although it is hard to decouple uniquely the contribution of each interlinked variables of the compact model, it is feasible to study the impact of some independent parameters by running different simulations changing their value. The impact of the thermal capacitance C_{th} and thermal resistance R_{th} is investigated in the following.
Since the I-V sweep is a quasi-static simulation and the discrete time intervals between 2 consecutive bias (V_A^i and V_A^{i+1} in Fig.3.6) are of the order of $\Delta t \sim 0.1\ ms$, the thermal time constant must be evaluated in order to understand why C_{th} has no impact on this simulation.

$$\tau_{th} = C_{th} \cdot R_{th} = 2.13 \cdot 10^{-16}\ JK^{-1} \cdot 6.3795 \cdot 10^{5}\ KW^{-1} \approx 136\ ps$$

A small τ_{th} value means that the switching volume can be heated up/cooled down very quickly, in this case approximately in less than 1 ns: this means that when the simulated time interval Δt is much larger than τ_{th}, a change of C_{th} does not influence the simulation and equation 3.11 could be simplified in its steady-state form:

$$C_{th} \cdot \frac{dT}{dt} = I_e \cdot V_A - \left(\frac{T - T_0}{R_{th}}\right) \approx 0 \quad \rightarrow \quad T = T_0 + R_{th} \cdot V_A \cdot I_e$$

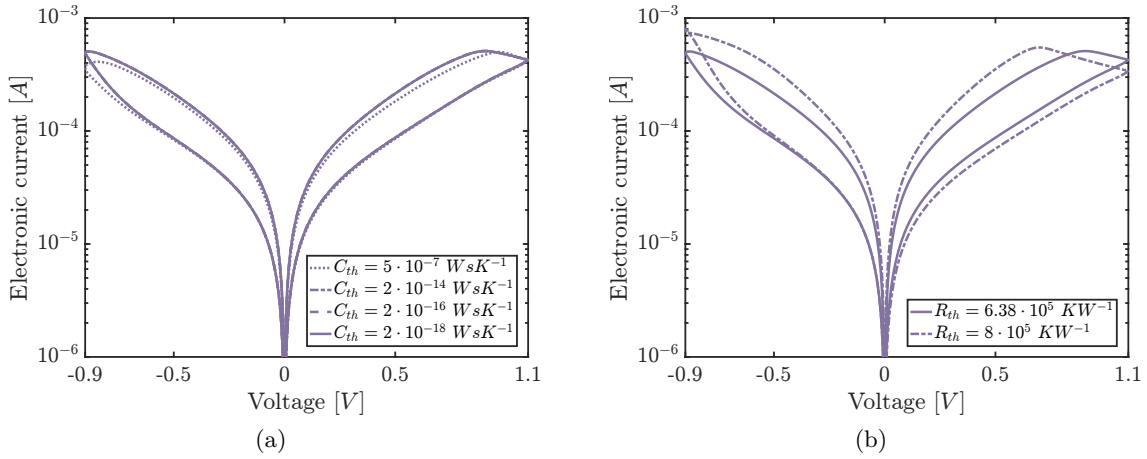


Figure 3.12: Impact of thermal model parameters on the I-V characteristic: (a) C_{th} and (b) R_{th}.

As shown in Fig.3.12a, thermal capacitances 2 orders of magnitude larger or smaller return overlapped I-V curves ($C_{th} = 2 \cdot \{10^{-14}; 10^{-16}; 10^{-18}\}\ JK^{-1}$). Only a very large C_{th} can impact on the simulation: with $C_{th} = 5 \cdot 10^{-7}$ the thermal time constant would be comparable to the simulated time intervals. However this big value is physically unreasonable, as the mass of the heated volume would be bigger than the device.
Furthermore, the insignificant impact of C_{th} confirms the hypothesis made in subsection 3.1.2: "to reduce the computational cost of the simulation, a unique value of c_p is chosen and it is computed as the average in the temperature range in which this device can work". Even if the temperature dependence of the specific heat had been included, the results would be concealed.
Conversely, R_{th} has a strong influence on the simulation, even if slightly changed: R_{th} indicates how much temperature difference is generated per unit electric power. The more is the

temperature increase (larger R_{th}), the wider is the resistance window (see Fig.3.12b) due to the fact that ion migration is enhanced at higher temperatures [35].

Multi-state modeling

As already mentioned in chapter 1, CMO/HfO_x analog ReRAM can be programmed in IRSs and there are two approaches to access them.

- Train of identical square voltage pulses can be sent to device to reproduce the potentiation and depression characteristic (see Fig.2.10) [41].
- Multiple quasi-static I-V sweeps with incremental current compliance or incremental stop voltage $|V_{Stop}|$ allow to access IRSs within the HRS/LRS window [39].

The algorithm illustrated in Fig.3.6 is repeated for multiple SET and multiple RESET to access IRSs by tuning the V_{Stop}^{SET} after each sweep. The simulated programming scheme employed to reproduce the I-V characteristic of 8 IRSs is shown in Fig.3.13. V_{Stop}^{SET} is decreased by 225 mV and V_{Stop}^{RESET} is increased by 288 mV between consecutive SET/RESET sweeps. The same sweep rate used in binary state simulation is choosen to reproduce Fig.3.13, i.e. SR = 0.1 Vs^{-1}.

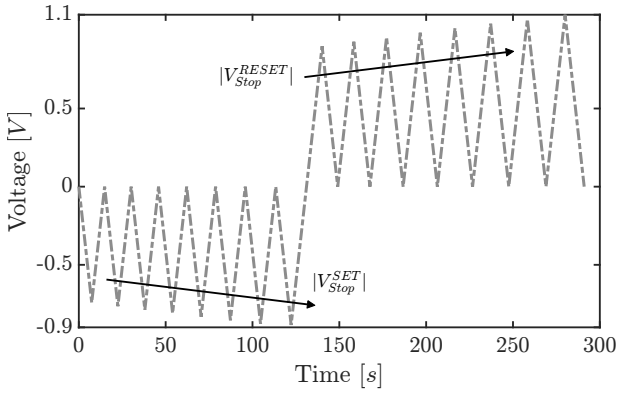


Figure 3.13: Simulated programming scheme to access IRSs with quasi-static multiple sweeps.

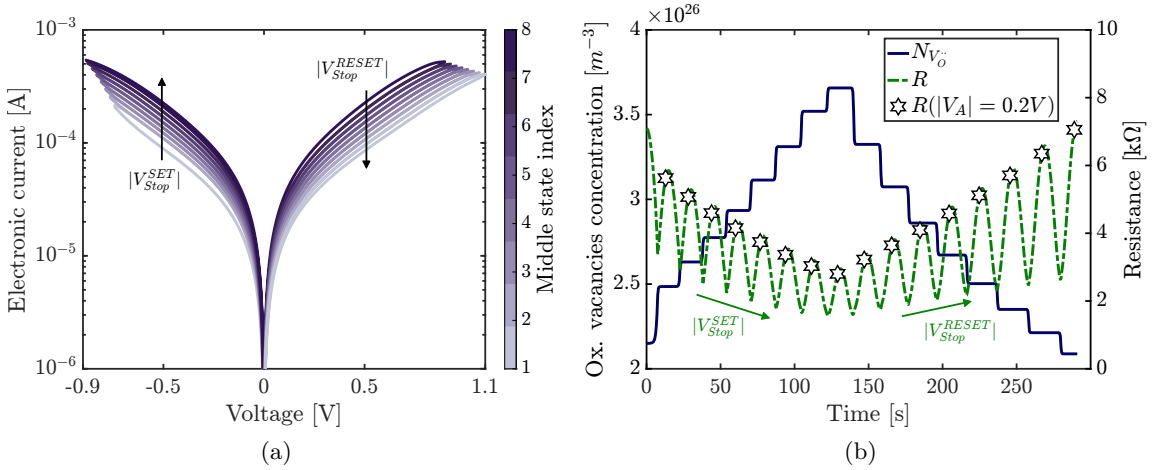


Figure 3.14: (a) Simulated clockwise I-V characteristic of TaO_x/HfO_x-based ReRAM device for 8 IRSs. (b) Evolution of oxygen vacancies concentration and total resistance of the device over the time interval of multiple sweeps.

The quasi-static I-V sweep extracted from the multi-state simulation (see Fig.3.14a) clearly shows the progressive change of the device conductance whenever a new sweep with larger $|V_{Stop}|$ is run.
Additionally, the superimposition of oxygen vacancies modulation (blue curve) and the total resistance of the device (green curve) in Fig.3.14b reflect the theoretical framework behind the model: different oxygen-deficient phases in the TaO_x alter the conduction properties of the device. A low bias (when no switching processes occur) is used to evaluate what happened after each single sweep. White stars in Fig.3.14b represent the total resistance when $|V_A| = 0.2\ V$ during each backward phase of the sweeps (after switching process). The monotonic decrease (SET) or increase (RESET) of the resistance at low bias demonstrate the analog properties of the device and the validity of the physical interpretation of IRSs, in fact $R(|V_A| = 0.2\ V)$ are aligned with distinct levels of $N_{V_O^{\cdot\cdot}}$.

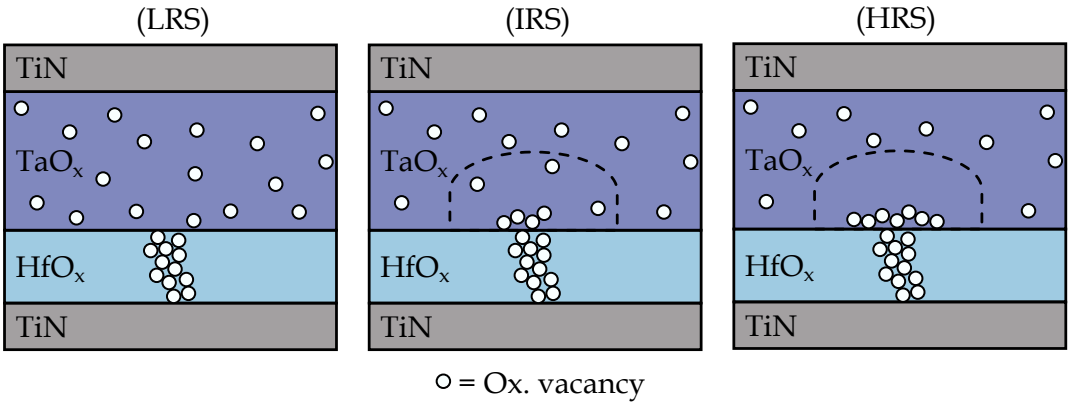


Figure 3.15: Oxygen vacancies spatial arrangement interpretation in TaO_x/HfOx ReRAM in LRS/IRS/HRS.

3.2.2 Pulse response model

To achieve real in-memory and analog neuromorphic computational tasks, ReRAM devices are programmed in the desired conductance state with multiple square pulses [44, 73, 74, 75]. It has been demonstrated that ultra-fast square pulses up to $\sim 10^2\ ps$ are suitable to induce resistive switching in TaO_x-based ReRAM [46, 76, 77], rather than using long triangular sweeps. For these reasons, the development of a compact model turned out to be crucial to accomplish IC simulations with CMO/HfO_x ReRAM techonology.
Here the compact model for TaO_x/HfO_x ReRAM device is extended to pulse response and partially validated with experimental data of single pulse experiment.
The algorithm used to simulate a square voltage pulse as input of the simulation is illustrated in Fig.3.7.

Single pulse

In order to reproduce a simulation of the single pulse experiment reported in subsection 2.1.2, a realistic pulse shape is designed with discrete rise time of $\Delta t_{rise} = 20\ ns$ to reach the voltage amplitude of square pulses. The pulse sequence (Fig.3.16a) is the same as in the experiment: pre- and post-pulse READ allow to check that the resistance switches between the desired HRS/LRS, while a long ($\sim \mu s$) single square pulse is employed to program the device conductance state. The chosen amplitude of READ pulses is $V_{pulse}^{READ} = +0.2\ V$, such small as to ensure no switching during the READ procedures. Electrical RC transients are not modeled in this simulation.

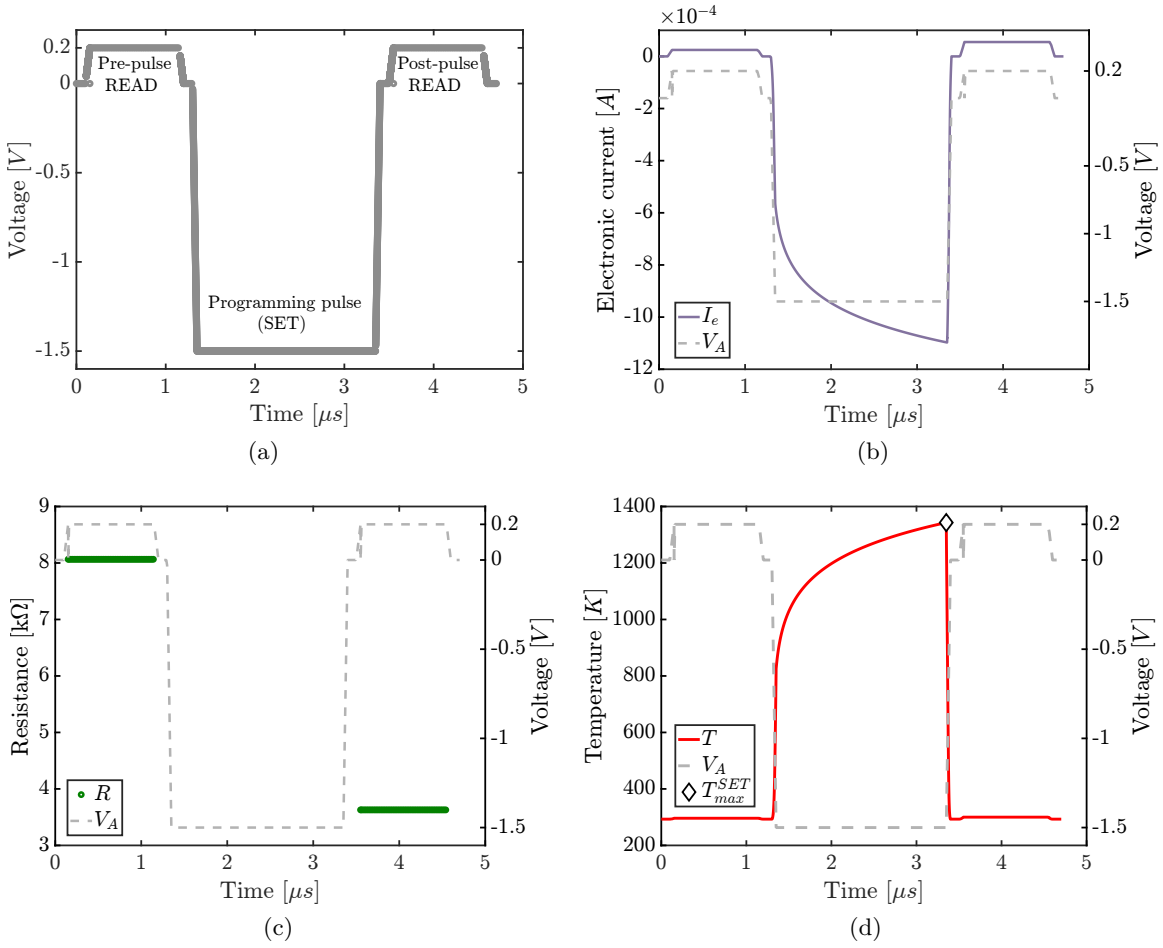


Figure 3.16: (a) Simulated pulse sequence READ-Programming SET pulse-READ. (b) $I(t)$ evolution as a response to the SET pulse sequence. (c) Resistance plotted during the time intervals of pre- and post-pulse READ. (d) Temperature evolution during the time interval of the SET pulse sequence.

Fig.3.16b shows the typical SET $I(t)$ characteristic extracted as response to single pulse switching: an increase of the current in absolute terms shows that negative programming pulses $V_{pulse} < 0$ allow to SET the device, as also demonstrated by the resistance drop from pre- to post-pulse READ times (Fig.3.16c). Thermal runaway phenomena, typical of SET transitions, lead to persistent T increase reaching a maximum temperature $T_{max}^{SET} > 1000\ K$, as shown in Fig.3.16d. T_{max}^{SET} in single pulse simulation are much larger if compared with the quasi-static simulation temperatures: the reason rests on the applied voltages involved in pulse mode, which are typically larger than quasi-static ones, enhancing electronic transport and Joule heating.

From Fig.3.17a to Fig.3.17d, the same simulation is repeated for the the RESET, i.e. a programming pulse with $V_{pulse} > 0$. The temperature plot in Fig.3.17d, shows that T_{max}^{RESET} is reached as soon as the V_{pulse}^{RESET} is applied: as already discussed for the quasi-static simulation results, the temperature transitions are instantaneous because the thermal time constant $\tau_{th} = C_{th} \cdot R_{th}$ is negligible with respect to the time intervals used for the sampling of the pulse sequence.

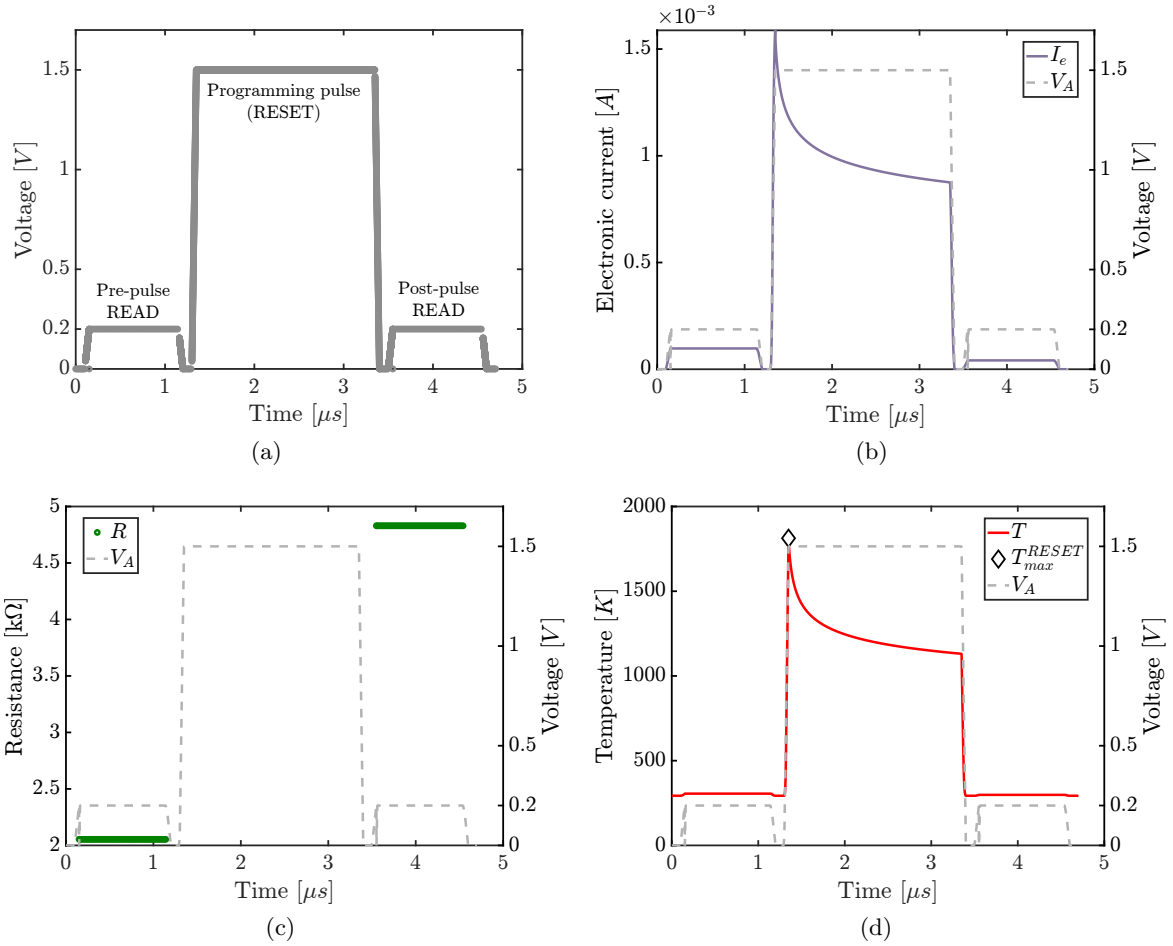


Figure 3.17: (a) Simulated pulse sequence READ-Programming RESET pulse-READ. (b) $I(t)$ evolution as a response to the RESET pulse sequence. (c) Resistance plotted during the time intervals of pre- and post-pulse READ. (d) Temperature evolution during the time interval of the RESET pulse sequence.

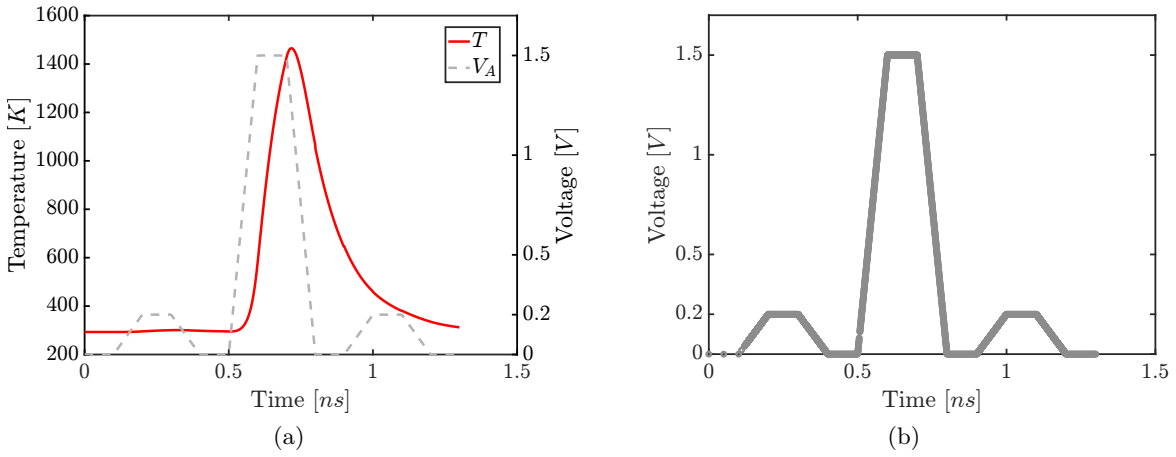


Figure 3.18: (a) Temperature evolution during the time interval of an ultra-short RESET pulse. (b) Simulated ultra-short pulse sequence READ-Programming RESET pulse-READ.

According to these insights, in an ultra-short pulse simulation the temperature transients to reach T_{max}^{RESET} should be visible: in agreement with these expectations, Fig.3.18a shows that in a simulation with raise time and pulse duration comparable with the thermal time constant, the temperature increase is not instantaneous. The pulse sequence in Fig.3.18b has the following features $\Delta t_{pulse}^{READ} = \Delta t_{pulse}^{RESET} = 100\ ps$ (shorter than $\tau_{th} = 136\ ps$).

In Fig3.16c the device is SET with a narrower resistance window with respect to the initial and final states (8 $k\Omega \rightarrow 2\ k\Omega$) chosen to characterize the t_{SET} in section 2.1. Since the goal is to develop an experimentally validated compact model for the device pulse response, the same pulse sequences used in the electrical characterizations are employed in the simulation. The conditions to be verified concern the resistance during the pre- and post-pulse READ:

- 8 $k\Omega \pm 100\ \Omega \rightarrow 2\ k\Omega \pm 100\ \Omega$ for the SET (see Fig.3.19a)
- 2 $k\Omega \pm 100\ \Omega \rightarrow 8\ k\Omega \pm 100\ \Omega$ for the RESET (see Fig.3.19b)

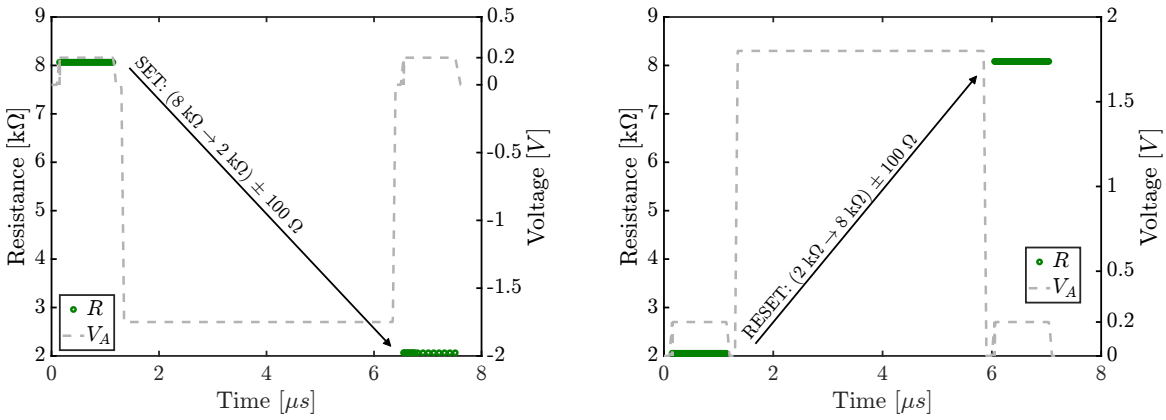


Figure 3.19: SET (a) and RESET (b) conditions to consider the pulse features as data for the Voltage-Time Trade-Off plot.

Plotting the experimental t_{SET} against the pulse amplitude V_{pulse}^{SET} reproduces the SET VTTO plot (Fig.3.20). The VTTO plot demonstrates the non-linear and exponential relation between the switching time and the amplitude of the programming/writing pulse in TaO_x/HfO_x ReRAM devices. This is consistent with findings reported in literature for other VCM devices [46, 70]. Regarding the VTTO plot in Fig.3.20, the measurement limit is determined by the RC charging time affecting the measurements, as detailed in section 2.1.2.

The experimental VTTO plot is used to validate the compact model for the device pulse response. A key finding of this study is that SET VTTO plots regarding simulated and experimental data overlap with high accuracy, as demonstrated with Fig.3.21a and Fig.3.21b. The simulation confirms the exponential trend (linear in logarithmic scale) of the time required to SET the device (t_{SET}) from 8 $k\Omega$ to 2 $k\Omega$ as a function of the pulse amplitude (V_{SET}). The simulated t_{SET} values are extracted using a trial and error approach, repeating the simulation until the resistance window condition is met while keeping V_{SET} fixed.

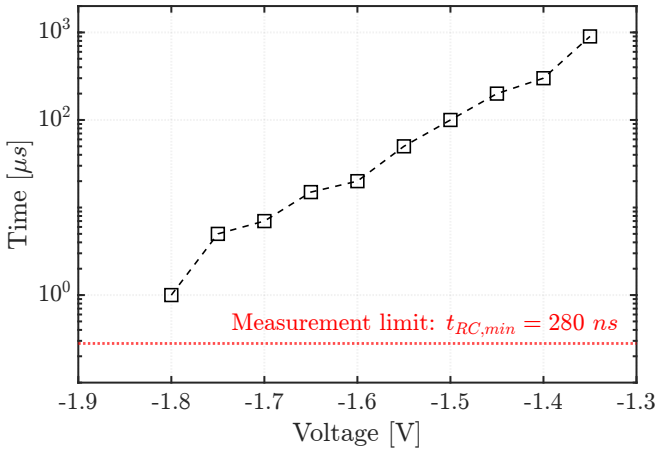


Figure 3.20: Experimental Voltage-Time Trade-Off plot showing the exponential $t_{SET}(V_{pulse}^{SET})$ relation in logarithmic scale.

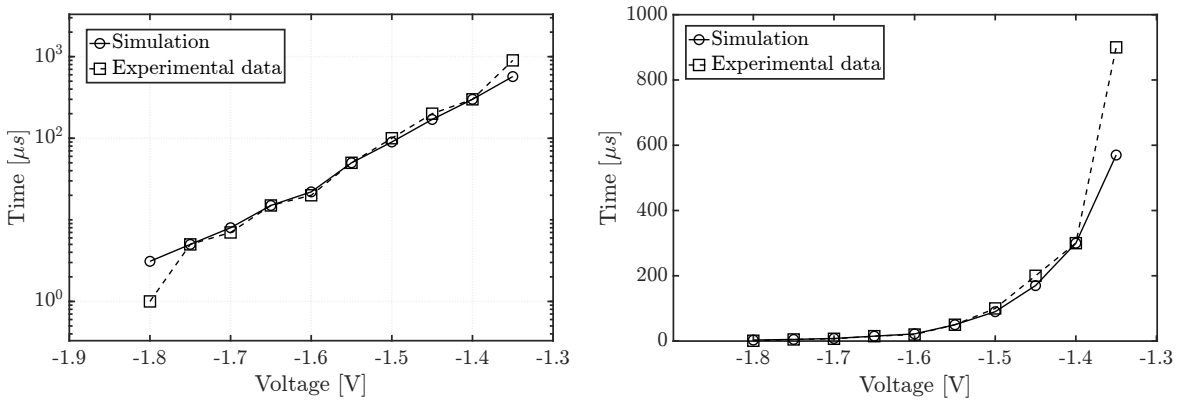


Figure 3.21: SET Voltage-Time Trade-Off plots with simulated and experimental data in logarithmic (a) and linear (b) scale.

Despite the experimental data are available only for the SET VTTO plot, the RESET one can be built with simulation data only, as the analysis conducted so far revealed promising and predictive results. Furthermore, simulation data of SET/RESET VTTO plots (Fig.3.22a and Fig.3.22b respectively) are fitted with an exponential law to extract the analytic relations $t_{SET} = f(V_{SET})$ and $t_{RESET} = f(V_{RESET})$ to perform 8 $k\Omega \rightarrow$ 2 $k\Omega$ and viceversa.

The fitting curve used in Fig.3.22 is in the form

$$t_{pulse} = t_0 \cdot \exp\left(\gamma \cdot V_{pulse}\right)$$

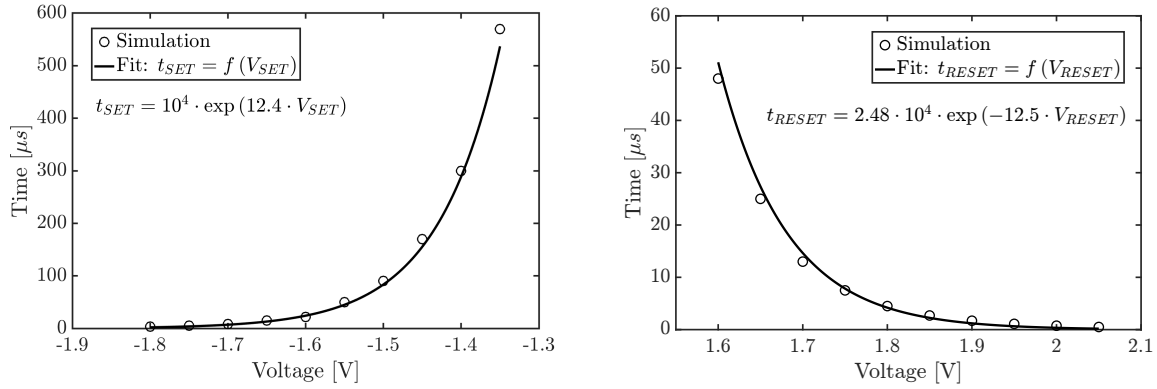
leading to the following relations:

$$t_{SET}^{(8\rightarrow 2)k\Omega} = 10^4 \cdot \exp\left(12.4 \cdot V_{SET}\right) \tag{3.12}$$

$$t_{RESET}^{(2\rightarrow 8)k\Omega} = 2.48 \cdot 10^4 \cdot \exp\left(-12.5 \cdot V_{RESET}\right) \tag{3.13}$$

However, it is not possible to state that the analytic relations 3.12 and 3.13 apply in general for SET/RESET operations with a single square pulse. They might be valid only

within the range of the considered pulse amplitudes and for the specific transitions involving 8 $k\Omega$ and 2 $k\Omega$ as HRS and LRS, respectively.



(a) (b)

Figure 3.22: Simulated SET (a) and RESET (b) Voltage-Time Trade-Off plot data fitted with exponential laws to extract the analytic $t_{pulse}(V_{pulse})$ relation.

As expected, the RESET time is less dependent on the amplitude of the pulse comparing it with the SET one: SET times characterized with single pulse switching are known to be highly non-linear in VCM [70]. Thermal runaway phenomena caused by the positive feedback between temperature and current makes the SET usually non-linear, whereas the RESET is not affected by this problem.
Even if the device studied in this work exhibits remarkable improvements concerning the SET non-linearity, the VTTO plots (Fig.3.22a and Fig.3.22b) still show that SET time (t_{SET}) follows an exponential dependence on V_{SET}, which is more pronounced than in the RESET case.

Pulse stream

The memristive compact model describing the response of TaO_x/HfO_x ReRAM operating in pulse mode can be used to show the noteworthy analog features of the device demonstrated in [41].
Bidirectional accumulating response characterizations depicted in Fig.2.10b reveal the presence of many IRSs, that, following the hypothesis of this model, are attributed to a partial modulation of defects concentration in the TaO_x dome (see Fig.3.15).
In subsection 3.2.1 it has been demonstrated that the model enables the description of IRSs as gradual modulation of the state variable $N_{V_O^{\cdot\cdot}}$. Furthermore, experimental validation of the single-pulse response suggests that it can be extended to a pulse stream employing the same algorithm repeated for n-pulses. Therefore a train of 20 square programming (10 up and 10 down) and 20 READ pulses is designed (Fig.3.23) to check the accumulative conductance response.
The pulse stream is applied to the device starting from a $125\mu S$ (8 $k\Omega$) state. As shown in Fig.3.23b, the current flowing in the device gradually increases over the time interval corresponding to negative ($V_A < 0$) pulses (potentiation), while positive (V_A) pulses causes a current decrease (depression). Fig.3.23c and Fig.3.23d confirm that the model is able to reproduce the analog bidirectional behavior of the device undergoing a train of up and down pulses.
Since the simulation of a potentiation/depression characteristic can be computationally expensive, the goal of the 10 up - 10 down test was to check that the device model works

properly.

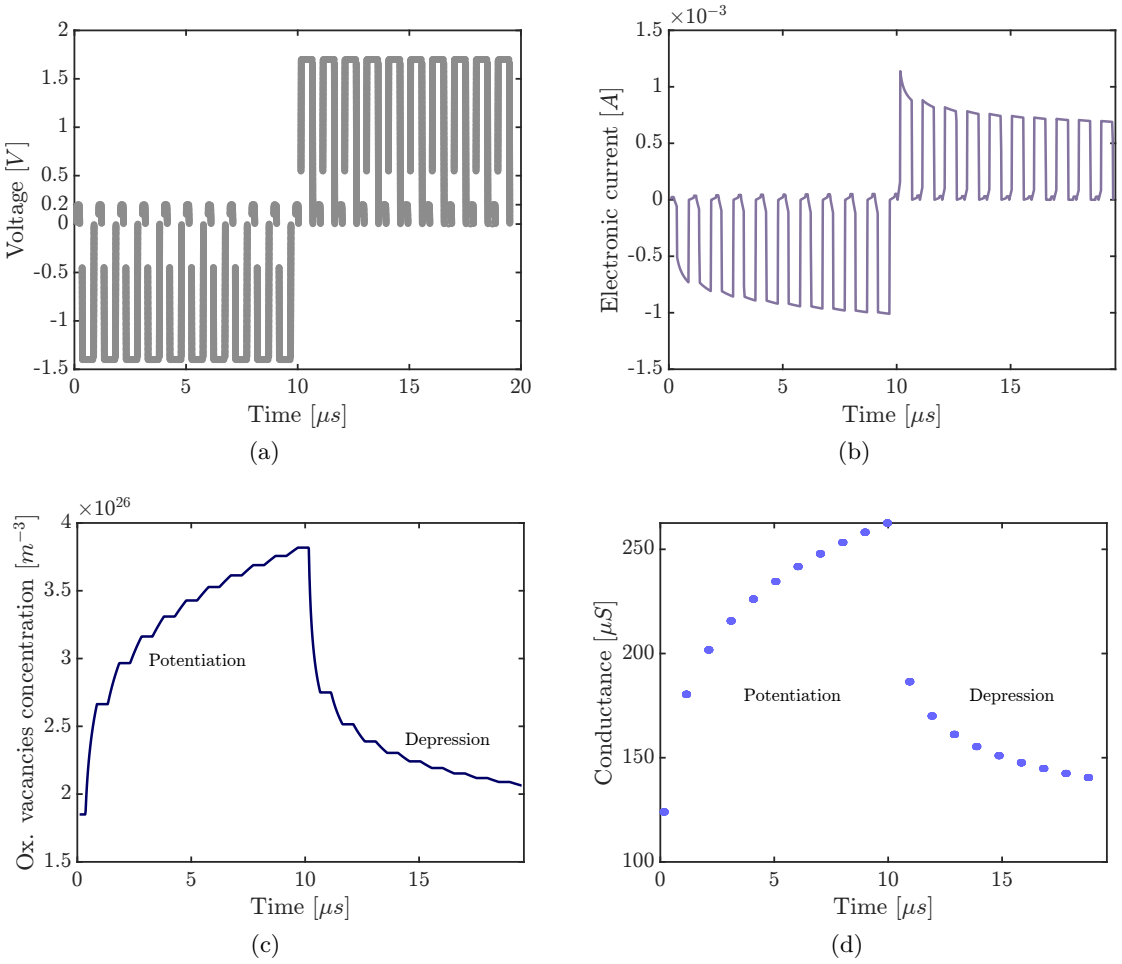


Figure 3.23: (a) Simulated programming scheme with 10 up-10 down pulses and 20 READ pulses. (b) $I(t)$ characteristic during the pulse stream time span. (c) Accumulative oxygen vacancy modulation. (d) Bidirectional potentiation/depression characteristic.

The second test consists in using the same number/width/duration of the pulses of the experimental potentiation/depression plot as input for the simulation to build the same pulse stream (see Fig.2.10a). Consequently, the simulated accumulative response of 200 up - 200 down pulses with the following features is compared with the experimental data:

- $\Delta t_{raise} = 50\ ns$
- $\Delta t_{pulse}^{READ} = 100\ ns$
- $\Delta t_{pulse}^{up} = \Delta t_{pulse}^{up} = 200\ ns$
- $V_{READ} = +0.2\ V$
- $V_{pulse}^{up} = +1.75\ V$
- $V_{pulse}^{down} = -1.25\ V$
- $n_{pulse}^{up} = n_{pulse}^{down} = 200$

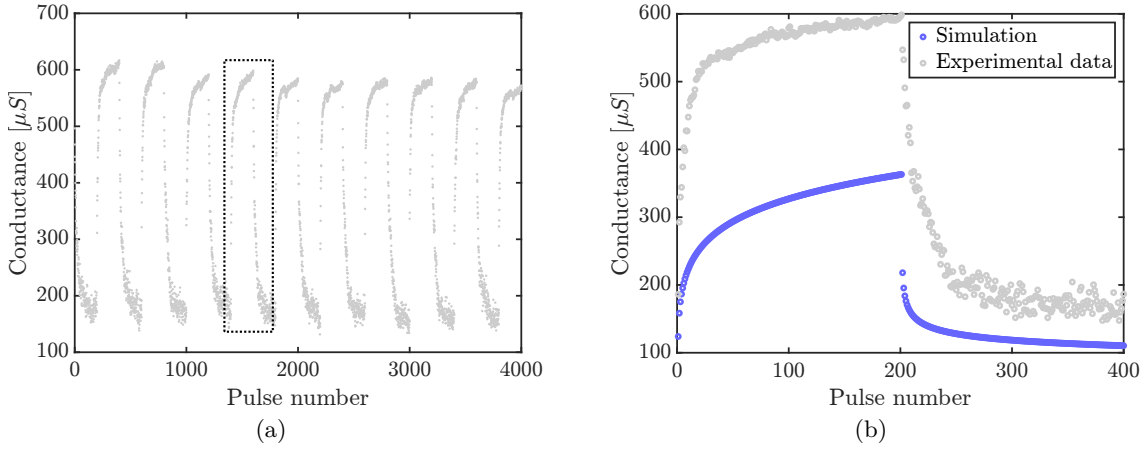


Figure 3.24: (a) 10 cycles of experimental potentiation/depression characteristic. (b) Comparison between simulated and experimental accumulative response using the same programming pulse scheme.

Contrary to previous findings, the pulse stream model is not matching perfectly the experimental results: in Fig.3.24b it is evident that the simulated conductance window is narrower than the experimental one. As shown in Fig.3.25, the accumulative response simulation almost overlaps the experimental one if multiplied by a factor 1.7x.

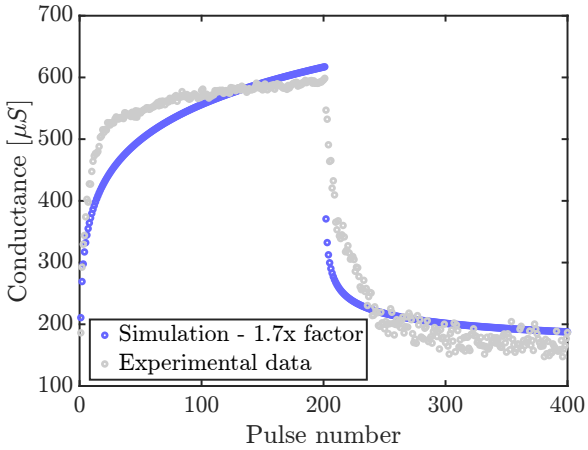


Figure 3.25: Simulated accumulative response corrected empirically to match the experimental conductance window.

It is hard to allocate this problem to physical explanations, since by definition, compact models might miss complex dynamic mechanisms that are not included to keep the computational cost of the simulation as low as possible and accurate at the same time. However, the problem of narrow conductance window could be merely solved with a re-calibration of model parameters, monitoring their compliance within physical limits.
Once refined, these outcomes might have significant implications for future research in the development of circuits based on analog CMO_x/HfO_x ReRAM device, enabling the simulation of computational tasks based on the ReRAM technology addressed in this work.

4 Conclusions

In this dissertation, a physics-based compact model of analog filamentary Conductive-Metal-Oxide/HfO_x ReRAM device is proposed. A substoichiometric TaO_x is chosen as case study of Conductive-Metal-Oxide. The goal of this compact model was to bridge the gap between the analog ReRAM IBM technology and its simulation for practical applications. Various aspects of the device have been investigated, ranging from the the underlying physical mechanisms to the electrical characterizations in both the quasi-static and AC domain. In particular, the understanding of the device physics turned out to be essential for the development of a realistic compact model able to catch the experimental data, laying the groundwork to predictive simulations involving this technology.

4.1 Key findings

Physical Mechanisms Oxygen vacancies migration is the key mechanisms used to describe the alteration of the conduction properties in the active layer of the device. The migration is modeled through hopping laws taking into account field- and temperature-driven processes. The resistance varies because of the different concentrations of electron trap states associated to defects, which limits the current flow. Conduction mechanisms are also coupled with a dynamic thermal model able to cover the temperature transients in the time domain. Combining these processes, the model is well suited to describe the SET and RESET dynamics. The intrinsically gradual RESET and the self-accelerated SET transitions are captured by the simulations and explained in terms of thermal runaway phenomena.

Device properties The physics-based framework allows the model to cover several aspects of the device behavior. In particular, the analog features of the device and the non-linearity of the switching mechanisms, are well reproduced by the simulations. Voltage pulse response simulations are employed to demonstrate the non-linear nature of the SET kinetics and the differences with the RESET one. The single pulse model is extended to simulate the switching behavior of the device when stimulated by voltage pulse stream. Although the high computational cost of the simulation, the analog bidirectional accumulative response of the conductance can be reproduced.

Model validation and accuracy Electrical characterization experimental data are used as a reference to validate the model in quasi-static and AC domain. The model covers with high accuracy the quasi-static single cycle I-V sweep data. In addition, the experimentally observed non-linear SET kinetics, derived from single pulse response characterizations, is precisely captured by the model. Additional corrections are needed to achieve the same accuracy also in accumulative response simulations.

Model classification This compact model falls in the branch of fully physics based simulations: all the parameters used in the simulation are checked to be physically reasonable and aligned with materials first principle-physics studies. Of particular significance is the coherence of the parameters used for simulations in different time domains: material/geometry/conduction parameters were never altered between quasi-static and transient models, while this could be required for other math-based/empirical compact models.

4.2 Future perspectives

Taken together, the findings suggest that the model can be employed to address circuit simulations for real applications. Despite the promising results represent a significant advancement to simulate TaO_x/HfO_x-based ReRAM, the model can be further improved.

Failure mechanisms modeling ReRAM devices suffer switching failure events that have a significant impact on the long-term performances of the device, such as data retention and endurance. Future research might be focused on understanding the physical mechanisms that lead the device to fail during the resistive switching operations.
Active materials where switching processes occur, might struggle to withstand mechanical/thermal stress caused by ion migration over either repeated cycling or pulsed switching. Additionally, high electric fields within the device could be critical, causing ion interdiffusion or dielectric breakdown in the worst case. Also excessive temperatures (as a result of too high current densities) damage the active materials, altering its switching properties. All these effects contribute to gradual or instantaneous degradation of the device performances, potentially leading to permanent failure.
Modeling failure mechanisms not only strengthens the understanding of the physical processes involved in the device but also provides insights and strategies to mitigate limitations for long-term reliability.

Noise model and stochasticity Delving the discussion into the potential improvements of this work, a noise model is still missing. Noise and stochasiticity are inherent characteristics of ReRAMs in general. Especially for the ReRAM technology investigated in this work, where conduction is trap-assisted, random fluctuation due to trapping/detrapping of carriers becomes particularly pertinent: as a result, stochastic variations of resistance level and current flow impact on reading/writing operations. This is usually referred to as Random Telegraph Noise (RTN). A variability-aware compact model including RTN effects can be developed taking as a reference statistically exhaustive reading/writing noise characterizations for the investigated device. It would enable realistic simulation of the ReRAM operation, predicting its behavior in integrated systems simulations despite non-ideal and random effects.

Modeling the 1T1R cell Analog CMO/HfO_x ReRAM devices are integrated in the BEOL of CMOS-based IC systems, thus a comprehensive compact model would include the impact of electrical elements in series with the resistive switching device. As the CMOS controlling element introduces parasitic resistances, voltage dividers between external resistors and the switching device have an impact on the electric field that contributes to switching phenomena in the ReRAM cell: this could potentially alter the switching window and the resistance levels. Moreover CMOS-integrated ReRAMs have to deal with parasitic capacitances, entailing additional charging transients that could influence the switching dynamics.
Incorporating the transistor parasitics in the model allows to describe the full cell (1T1R) configuration, splitting the external contributions within the simulations of the resistive switching

cell.

In conclusion, further research could be focused on the optimizations of the model, advancing towards realistic simulations of analog CMO/HfO_x ReRAM IBM technology in cutting-edge memory architectures.

A Appendix - Fabrication process of HfO_x/TaO_x ReRAM device

As introduced in section 1.3, the term "*bilayer*" means that the oxides dividing the TE and BE are 2 stacked on top of the other. In this case, the oxides bilayer is composed of a substoichiometric Hafnium Oxide ($HfO_{x<2}$) below (acting as dielectric layer) and a substoichiometric Tantalum Oxide (TaO_x) on top, whereas the metal electrodes are made of Titanium Nitride (TiN). TaO_x is a TMO deposited in such a way as to be more conductive than an usual insulator and the requirements about its electrical conductivity are discussed in chapter 3.

The ReRAM device is called "*vertical*" for two reasons:

- layers are stacked on top of each other
- the top metal surface and the bottom substrate refer respectively to the electrical GND/Top Terminal (TT).

Furthermore, it is worth to mention that all the materials and technological processes involved in the fabrication of this device are CMOS and BEOL compatible.

The fabrication process flow starts with a Silicon (Si) substrate heavily doped with Arsenic (n^{++}), such as to have a conductive bottom contact. In order to ensure a good quality of the bottom electrical terminal, the native oxide on top of n^{++}-Si substrate must be removed, since the TiN BE is deposited directly on top of the Si. Therefore a water solution of ammonium fluoride (NH_4F) and hydrofluoric acid (HF), also known as Buffered Oxide Etch (BOE), is used to remove the native silicon dioxide (SiO_x) by substrate immersion. Then a 20 *nm* thick TiN layer and a 3.5 *nm* thick HfO_x layer are deposited by Plasma-Enhanced Atomic Layer Deposition (PEALD).
To deposit the TiN, the Tetrakis(DiMethylAmido)Titanium (TDMAT) is chosen as metal precursor, while the second reactant is a nitrogen plasma.
Concerning the HfO_x, Tetrakis(EthylMethylAmino)Hafnium and oxygen plasma are the metal precursor and the reactant respectively. The deposition step for HfO_x follows the TiN one without braking the vacuum of the PEALD chamber, in order to prevent the oxidation of the TiN.
The nominal 20 *nm* of substoichiometric TaO_x are deposited by DC reactive magnetron sputtering, where Argon (Ar) ions coming from the Ar flow in the chamber are accelerated to hit the Ta target, removing Ta (sputtered) atoms from the surface. Ta ions interact with the oxygen flow and by controlling the pressure of the chamber, TaO_x is deposited with different stoichiometries. However, after the TaO_x deposition step, multiple Ta-O substoichiometries co-exist, rather than a unique substoichiometric phase [26].
20 *nm* of TiN and 50 *nm* of tungsten (W) are deposited sequentially by RF magnetron sputtering (TiN target) and DC magnetron sputtering (W target) respectively. Fig.A.1a shows the stack deposited up to this step.

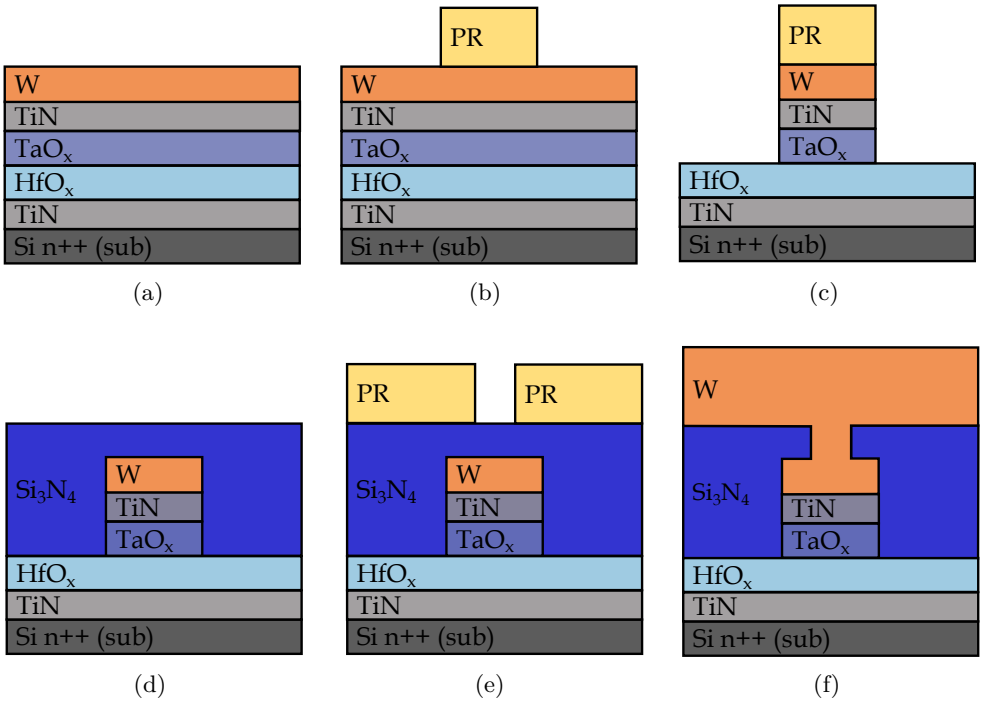
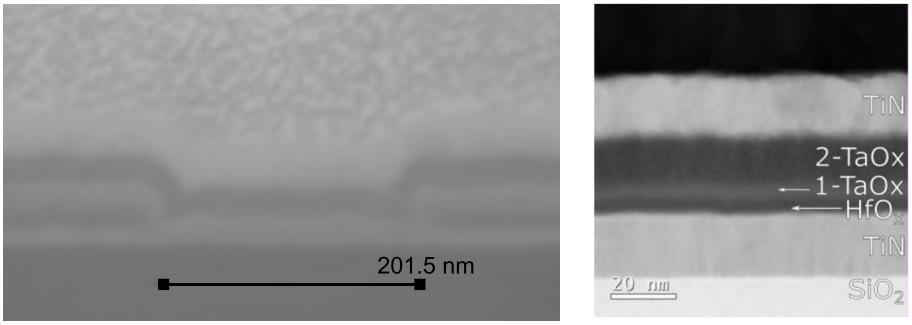


Figure A.1: Fabrication process flow of TiN-TaO_x-HfO_x-TiN ReRAM device in cross-section view: (a) Deposition of ReRAM material stack on a Si substrate. (b) Photoresist lithography patterning to define the cell area. (c) Plasma etching. (d) Si_3N_4 passivation. (e) Photoresist lithography patterning to define the via area. (f) Deposition of W top electrode. Materials geometry is not in scale.

Spinning and laser developing steps of positive Photoresist (PR) follow, such as to define the (200 $nm \times 200$ nm) cell area, called "*mesa*" (Fig.A.1b). The area outside the defined mesa is etched through Inductively Coupled Plasma Reactive-Ion Etching (ICP RIE) with a mixture of trifluoromethane (CHF_3), nitrogen (N_2) and sulfur hexafluoride (SF_6). The etching process is HfO_x-selective, hence only W, TiN and TaO_x are etched (Fig.A.1c).
After PR stripping, a Si_3N_4 passivation layer (cladding) is deposited by Plasma-Enhanced Chemical Vapor Deposition (PECVD), introducing in the PECVD chamber N_2 and silane (SiH_4) plasma (Fig.A.1d). The absence of oxygen species in the cladding material prevents further oxidation of the TaO_x layer.
Then a positive PR layer is spun and developed by direct laser writing in order to define the via for the TE access (Fig.A.1e): RIE with a O_2/CHF_3 gas is employed to etch the Si_3N_4 cladding in the via area. Finally, the PR is dissolved and the W TE is deposited by DC magnetron sputtering, resulting in the device whose schematic cross-section is depicted in FigA.1f.

Fig.A.2a and Fig.A.2b show the Scanning Electron Microscope (SEM) and Bright field Scanning Transmission Electron Microscope (STEM) images for the fabricated ReRAM device cross-section. The SEM image demonstrates that the expected mesa dimension is effectively 200 nm, while the STEM image reveals the presence of different materials (different contrast gradations).
Moreover, the STEM image shows that the TaO_x layer is more oxidized at the interface with the HfO_x, so in practice, it has 2 phases: ~ 3 nm of amorphous (1-TaO_x in the image) layer, which is more oxidized than the other ~ 17 nm (2-TaO_x in the image).



(a) (b)

Figure A.2: (a) Scanning Electron Miscroscope image of the ReRAM device stack cross-section [41], highlighting the 200 nm width of the cell. (b) Bright field Scanning Transmission Electron Microscope image of the ReRAM device stack cross-section [42].

B Appendix - Switching time characterizations

In the following appendix, there are the plots about the $I(t)$ evolution as a response to the experimental SET pulse sequence depicted in Fig.2.8a. 10 square voltage pulses are employed in the experiment, with incremental amplitude: $V_{pulse}^{SET} = \{-1.35 \div -1.8\}\ V$.

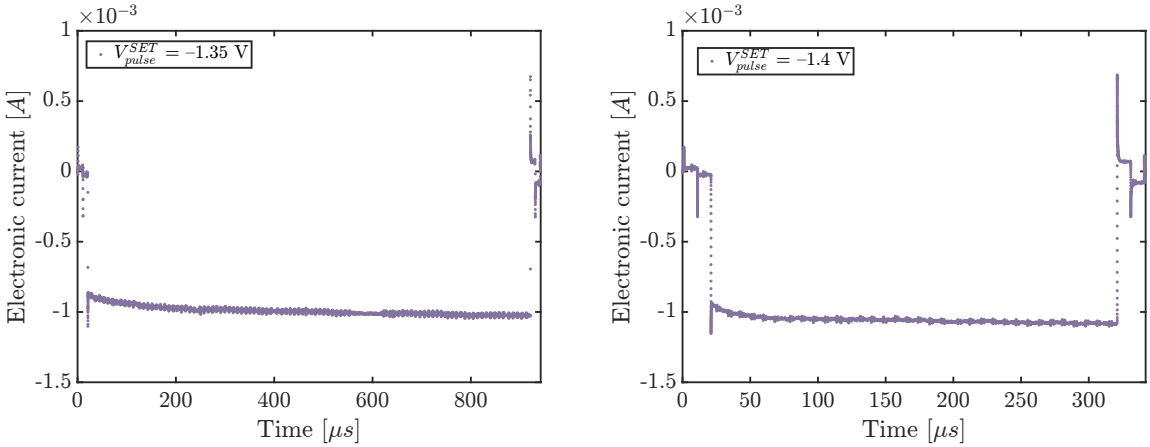


Figure B.1: $I(t)$ evolution in the switching time characterization.

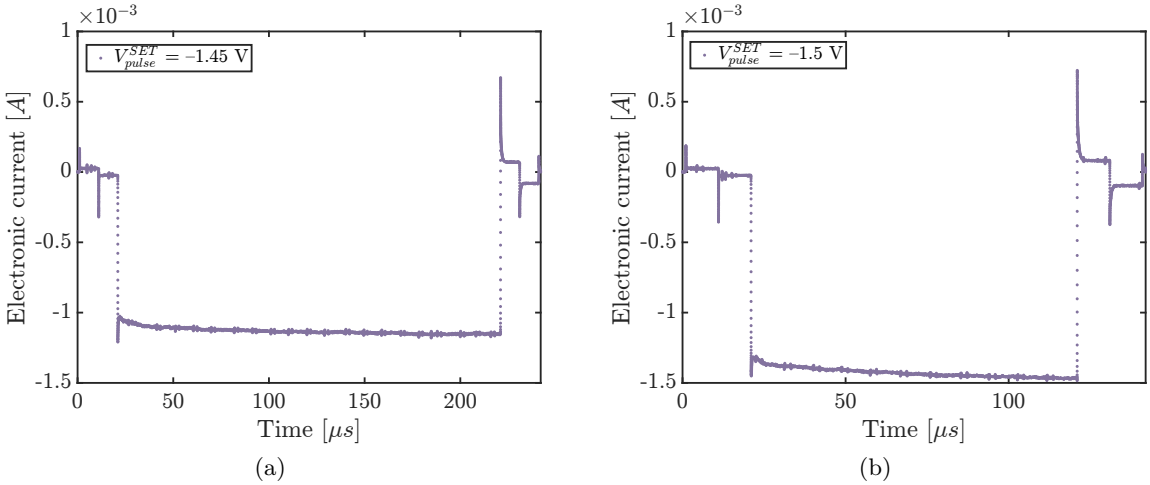


Figure B.2: $I(t)$ evolution in the switching time characterization.

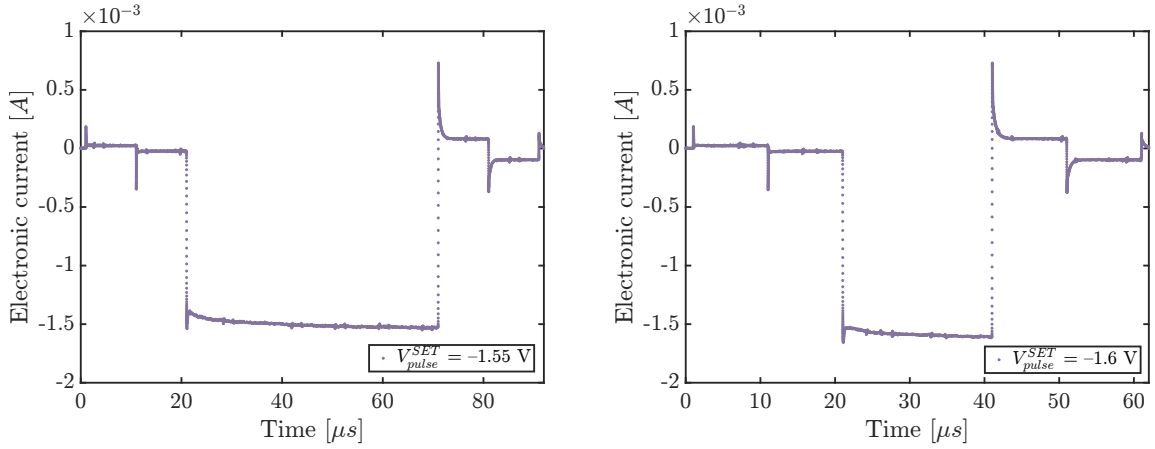


Figure B.3: $I(t)$ evolution in the switching time characterization.

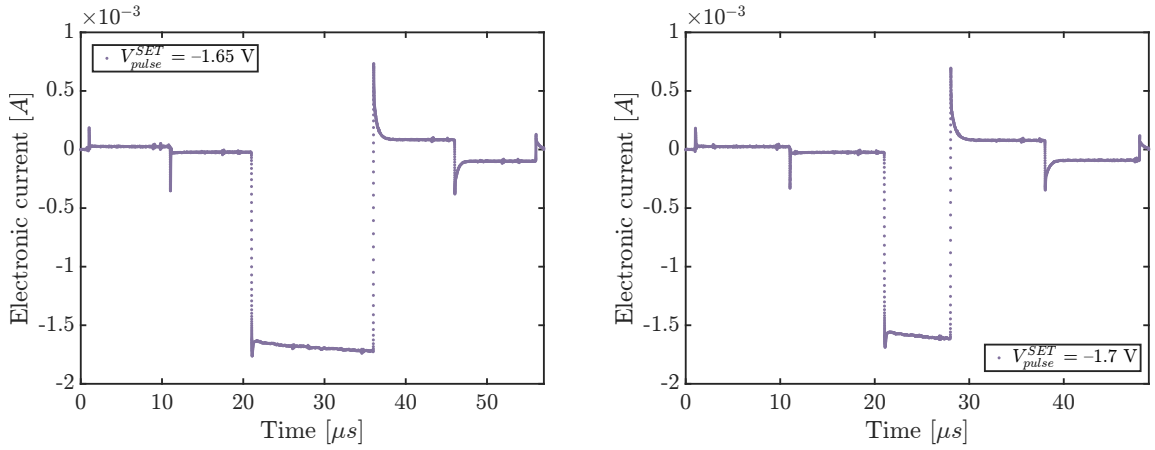


Figure B.4: $I(t)$ evolution in the switching time characterization.

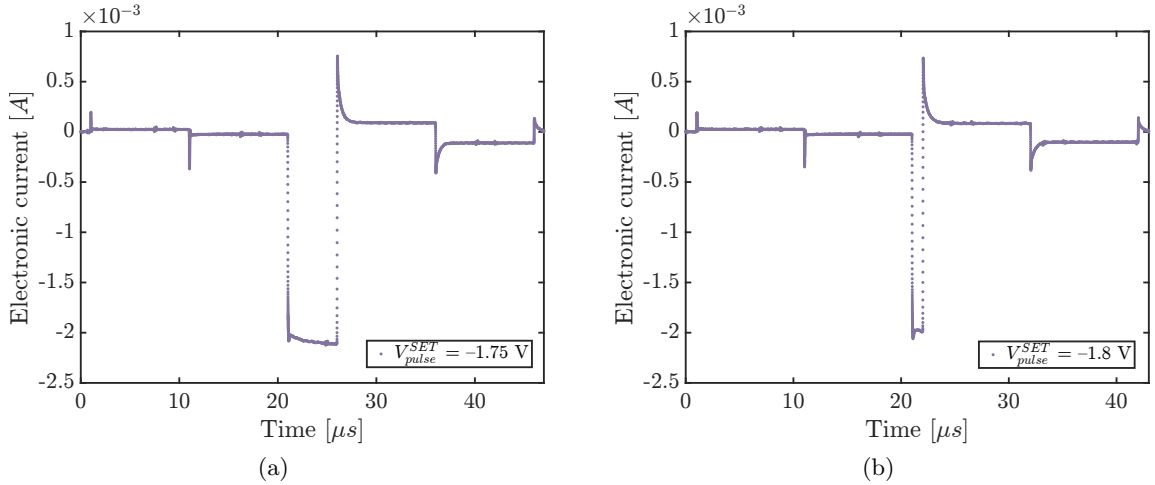


Figure B.5: $I(t)$ evolution in the switching time characterization.

C Appendix - Matrix formalism to solve the non-linear system

According to the mathematical conventions used in section 2.3, the equation system including 3.2, 3.3, 3.8 and 3.11 is written in the matrix formalism as follow:

$$\bar{\mathcal{X}} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} N^i_{\mathrm{V}^{\cdot\cdot}_{\mathrm{O}}} \\ I^i_{ion} \\ I^i_e \\ T^i \end{bmatrix}$$

$$\bar{\mathcal{F}} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix} = \begin{bmatrix} N^i_{\mathrm{V}^{\cdot\cdot}_{\mathrm{O}}} - N^{i-1}_{\mathrm{V}^{\cdot\cdot}_{\mathrm{O}}} + \Delta t \cdot \left(\frac{1}{qzV_{dome}}\right) \cdot \frac{\left(I^i_{ion} + I^{i-1}_{ion}\right)}{2} \\ I^i_{ion} - A_{dome} \cdot zqN^i_{\mathrm{V}^{\cdot\cdot}_{\mathrm{O}}} a\nu_0 \cdot \exp\left(-\frac{\Delta W_A}{k_B T^i}\right) \cdot 2\sinh\left(\frac{zq\mathcal{E}a}{2k_B T^i}\right) \\ I^i_e - A_{dome} q\beta z N^i_{\mathrm{V}^{\cdot\cdot}_{\mathrm{O}}} a_e \nu_e \cdot \exp\left(-\frac{\Delta E_A}{k_B T^i}\right) \cdot 2\sinh\left(\frac{q\mathcal{E}a_e}{2k_B T^i}\right) \\ T^i - T^{i-1} - \frac{\Delta t}{C_{th}} \cdot \left[I^i_e \cdot V_A - \left(\frac{T^i - T_0}{R_{th}}\right)\right] \end{bmatrix}$$

$$\mathcal{J} = \begin{bmatrix} \frac{\partial f_1}{\partial N_{\mathrm{V}^{\cdot\cdot}_{\mathrm{O}}}} & \frac{\partial f_1}{\partial I_{ion}} & \frac{\partial f_1}{\partial I_e} & \frac{\partial f_1}{\partial T} \\ \frac{\partial f_2}{\partial N_{\mathrm{V}^{\cdot\cdot}_{\mathrm{O}}}} & \frac{\partial f_2}{\partial I_{ion}} & \frac{\partial f_2}{\partial I_e} & \frac{\partial f_2}{\partial T} \\ \frac{\partial f_3}{\partial N_{\mathrm{V}^{\cdot\cdot}_{\mathrm{O}}}} & \frac{\partial f_3}{\partial I_{ion}} & \frac{\partial f_3}{\partial I_e} & \frac{\partial f_3}{\partial T} \\ \frac{\partial f_4}{\partial N_{\mathrm{V}^{\cdot\cdot}_{\mathrm{O}}}} & \frac{\partial f_4}{\partial I_{ion}} & \frac{\partial f_4}{\partial I_e} & \frac{\partial f_4}{\partial T} \end{bmatrix} = \begin{bmatrix} 1 & \frac{\Delta t}{2qzV_{dome}} & 0 & 0 \\ \frac{\partial f_2}{\partial N_{\mathrm{V}^{\cdot\cdot}_{\mathrm{O}}}} & 1 & 0 & \frac{\partial f_2}{\partial T} \\ \frac{\partial f_3}{\partial N_{\mathrm{V}^{\cdot\cdot}_{\mathrm{O}}}} & 0 & 1 & \frac{\partial f_3}{\partial T} \\ 0 & 0 & -\frac{\Delta t \cdot V_A}{C_{th}} & 1 + \frac{\Delta t}{C_{th} \cdot R_{th}} \end{bmatrix}$$

where the entries $\mathcal{J}_{2,1}$, $\mathcal{J}_{2,4}$, $\mathcal{J}_{3,1}$, $\mathcal{J}_{3,4}$ are:

$$\frac{\partial f_2}{\partial N_{\mathrm{V}_{\mathrm{O}}^{\cdot\cdot}}} = -A_{dome} \cdot zqa\nu_0 \cdot \exp\left(-\frac{\Delta W_A}{k_B T}\right) \cdot 2\sinh\left(\frac{zq\mathcal{E}a}{2k_B T}\right)$$

$$\frac{\partial f_2}{\partial T} = -2A_{dome} \cdot zqN_{\mathrm{V}_{\mathrm{O}}^{\cdot\cdot}} a\nu_0 \cdot \exp\left(-\frac{\Delta W_A}{k_B T}\right) \cdot \frac{1}{k_B T^2} \cdot \left[\Delta W_A \sinh\left(\frac{zq\mathcal{E}a}{2k_B T}\right) - \frac{zq\mathcal{E}a}{2} \cdot \cosh\left(\frac{zq\mathcal{E}a}{2k_B T}\right)\right]$$

$$\frac{\partial f_3}{\partial N_{\mathrm{V}_{\mathrm{O}}^{\cdot\cdot}}} = -A_{dome} \cdot zq\beta a_e \nu_e \cdot \exp\left(-\frac{\Delta E_A}{k_B T}\right) \cdot 2\sinh\left(\frac{q\mathcal{E}a_e}{2k_B T}\right)$$

$$\frac{\partial f_3}{\partial T} = -2A_{dome} \cdot zq\beta N_{\mathrm{V}_{\mathrm{O}}^{\cdot\cdot}} a_e \nu_e \cdot \exp\left(-\frac{\Delta E_A}{k_B T}\right) \cdot \frac{1}{k_B T^2} \cdot \left[\Delta E_A \sinh\left(\frac{q\mathcal{E}a_e}{2k_B T}\right) - \frac{q\mathcal{E}a_e}{2} \cdot \cosh\left(\frac{q\mathcal{E}a_e}{2k_B T}\right)\right]$$

Bibliography

[1] J. von Neumann, "First draft of a report on the EDVAC." In: IEEE Annals of the History of Computing 15.4 (1993), pp. 27–75.

[2] M. Suri, "Applications of Emerging Memory Technology." Springer Singapore, 2020.

[3] M.A. Zidan, J.P. Strachan and W.D. Lu, "The future of electronics based on memristive systems." Nature electronics 1.1 (2018): 22-29.

[4] J.L. Hennessy and D.A. Patterson, "Computer architecture: a quantitative approach." Elsevier, 2011.

[5] G.E. Moore, "Cramming more components onto integrated circuits." Proceedings of the IEEE 86.1 (1998): 82-85.

[6] K. Rupp, M. Horovitz and F. Labonte, "Years of microprocessor trend data." by karl-rupp.net (Online).

[7] J. Vidal, "Tsunami of Data Could Consume One Fifth of Global Electricity by 2025." (2017. Climate Home News)

[8] A. Mehonic and A.J. Kenyon, "Brain-inspired computing needs a master plan." Nature 604.7905 (2022): 255-260.

[9] D. Patterson et al., "A case for intelligent RAM." IEEE micro 17.2 (1997): 34-44.

[10] J. Kim and Y. Kim, "HBM: Memory solution for bandwidth-hungry processors." 2014 IEEE Hot Chips 26 Symposium (HCS). IEEE, 2014.

[11] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh and E. Eleftheriou, "Memory devices and applications for in-memory computing." Nature nanotechnology (2020), 15(7), 529-544.

[12] L. Chua, "Memristor-the missing circuit element." IEEE Transactions on circuit theory 18.5 (1971): 507-519.

[13] D.B. Strukov, G.S. Snider, D.R. Stewart and R.S. Williams, "The missing memristor found." Nature (2008), 453(7191), 80-83.

[14] I.H. Im, S.J. Kim and H.W. Jang, "Memristive devices for new computing paradigms." Advanced Intelligent Systems (2020), 2(11), 2000105.

[15] F. Zhang et al., "XMA: a crossbar-aware multi-task adaption framework via shift-based mask learning method." Proceedings of the 59th ACM/IEEE Design Automation Conference. 2022.

[16] T. Gokmen and Y. Vlasov, "Acceleration of deep neural network training with resistive cross-point devices: Design considerations." Frontiers in neuroscience 10 (2016): 203376.

[17] F. Aguirre et al., "Hardware implementation of memristor-based artificial neural networks." Nature Communications 15.1 (2024): 1974.

[18] Z. Sun et al., "Solving matrix equations in one step with cross-point resistive arrays." Proceedings of the National Academy of Sciences 116.10 (2019): 4123-4128.

[19] G.W. Burr et al., "Neuromorphic computing using non-volatile memory." Advances in Physics: X 2.1 (2017): 89-124.

[20] M. Lanza et al, "Memristive technologies for data storage, computation, encryption, and radio-frequency communication." Science 376.6597 (2022).

[21] Z. Wang et al., "Resistive switching materials for information processing." Nature Reviews Materials 5.3 (2020): 173-195.

[22] D. Ielmini, "Resistive switching memories based on metal oxides: mechanisms, reliability and scaling." Semiconductor Science and Technology 31.6 (2016): 063002.

[23] D.V. Christensen et al., "2022 roadmap on neuromorphic computing and engineering." Neuromorphic Computing and Engineering 2.2 (2022): 022501.

[24] F. Zahoor, T.Z.A. Zulkifli and F.A. Khanday, "Resistive random access memory (RRAM): an overview of materials, switching mechanism, performance, multilevel cell (MLC) storage, modeling, and applications." Nanoscale research letters (2020), 15, 1-26.

[25] N. Kanegami, Y. Nishi and T. Kimoto, "Unique resistive switching phenomena exhibiting both filament-type and interface-type switching in Ti/$Pr_{0.7}Ca_{0.3}MnO_{3-\delta}$/Pt ReRAM cells." Applied Physics Letters 116.1 (2020).

[26] T. Stecconi et al., "Filamentary TaO_x/HfO2 ReRAM Devices for Neural Networks Training with Analog In-Memory Computing." Advanced Electronic Materials 8.10 (2022).

[27] T. Stecconi et al., "Role of Conductive-Metal-Oxide to HfO_x Interfacial Layer on the Switching Properties of Bilayer TaO_x/HfO_x ReRAM." ESSDERC 2022-IEEE 52nd European Solid-State Device Research Conference (ESSDERC). IEEE, 2022.

[28] W. Wu et al., "Improving analog switching in HfO_x-based resistive memory with a thermal enhanced layer." IEEE Electron Device Letters 38.8 (2017): 1019-1022.

[29] S. Larentis, F. Nardi, S. Balatti, D.C. Gilmer and D. Ielmini, "Resistive switching by voltage-driven ion migration in bipolar RRAM—Part II: Modeling." IEEE Transactions on Electron Devices (2012), 59(9), 2468-2475.

[30] N. Gong et al., "Deep learning acceleration in 14nm CMOS compatible ReRAM array: device, material and algorithm co-optimization." 2022 International Electron Devices Meeting (IEDM). IEEE, 2022.

[31] R. Waser and M. Aono., "Nanoionics-based resistive switching memories." Nature materials 6.11 (2007): 833-840.

[32] S. Yin, X. Sun, S. Yu and J.S. Seo, "High-throughput in-memory computing for binary deep neural networks with monolithically integrated RRAM and 90-nm CMOS." IEEE Transactions on Electron Devices, 67(10), 4185-4192. (2020).

[33] Q. Liu et al., "33.2 A fully integrated analog ReRAM based 78.4 TOPS/W compute-in-memory chip with fully parallel MAC computing." 2020 IEEE International Solid-State Circuits Conference-(ISSCC).

[34] R. Mochida et al., "A 4M synapses integrated analog ReRAM based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture." 2018 IEEE Symposium on VLSI Technology.

[35] M. Lanza et al., "Recommended methods to study resistive switching devices." Advanced Electronic Materials 5.1 (2019): 1800143.

[36] E. Linn, A. Siemon, R. Waser and S. Menzel, "Applicability of well-established memristive models for simulations of resistive switching devices." IEEE Transactions on Circuits and Systems I: Regular Papers, 61(8), 2402-2410. (2014).

[37] A. Siemon et al., "Simulation of TaO_x-based complementary resistive switches by a physics-based memristive model." 2014 IEEE international symposium on circuits and systems (ISCAS).

[38] C. La Torre, A.F. Zurhelle and S. Menzel, "Compact modelling of resistive switching devices based on the valence change mechanism." International Conference on Simulation of Semiconductor Processes and Devices (SISPAD). IEEE, 2019.

[39] D.F. Falcone et al., "Analytical modelling of the transport in analog filamentary conductive-metal-oxide/HfO_x ReRAM devices." Nanoscale Horizons (2024).

[40] D.F. Falcone et al., "Physical modeling and design rules of analog Conductive Metal Oxide-HfO2 ReRAM." 2023 IEEE International Memory Workshop (IMW).

[41] T. Stecconi et al., "Analog Resistive Switching Devices for Training Deep Neural Networks with the Novel Tiki-Taka Algorithm." Nano Letters 24.3 (2024): 866-872.

[42] T. Stecconi et al., "Equivalent electrical circuit modelling of a TaO_x/HfO_x based RRAM with optimized resistance window and multilevel states." 2022 Device Research Conference (DRC). IEEE, 2022.

[43] L. Zhu, J. Zhou, Z. Guo and Z. Sun, "Synergistic resistive switching mechanism of oxygen vacancies and metal interstitials in Ta_2O_5." The Journal of Physical Chemistry C (2016), 120(4), 2456-2463.

[44] D. Ielmini and R. Waser, "Resistive switching: from fundamentals of nanoionic redox processes to memristive device applications." John Wiley & Sons, 2015.

[45] R. Waser, R. Dittmann, G. Staikov, and K. Szot, "Redox-based resistive switching memories -nanoionic mechanisms, prospects, and challenges." Advanced Materials (Deerfield Beach, Fla.) 21.25-26 (2009): 2632-2663.

[46] U. Böttger et al.,. "Picosecond multilevel resistive switching in tantalum oxide thin films." Scientific reports 10.1 (2020): 16391.

[47] J.J. Yang, D.B. Strukov and D.R Stewart, "Memristive devices for computing." Nature nanotechnology (2013), 8(1), 13-24.

[48] N.F. Mott and R.W. Gurney, "Electronic Processes in Ionic Crystals." (1940).

[49] S. Menzel, "Modeling and simulation of resistive switching devices." Doctoral dissertation, Aachen, Techn. Hochsch., 2012.

[50] M. Noman and W. Jiang, "Computational investigations into the operating window for memristive devices based on homogeneous ionic motion." Applied Physics A 102 (2011): 877-883.

[51] F.C. Chiu, "A review on conduction mechanisms in dielectric films." Advances in Materials Science and Engineering (2014).

[52] D. Schön and S. Menzel, "Spatio-Temporal Correlations in Memristive Crossbar Arrays due to Thermal Effects." Advanced functional materials (2023), 33(22), 2213943.

[53] T.D. Brown, S. Kumar and R.S. Williams, "Physics-based compact modeling of electrothermal memristors: Negative differential resistance, local activity, and non-local dynamical bifurcations." Applied Physics Reviews 9.1 (2022).

[54] S. Kumar, J.P. Strachan and R.S. Williams, "Chaotic dynamics in nanoscale NbO2 Mott memristors for analogue computing." Nature 548.7667 (2017): 318-321.

[55] H.J. Stetter, "Analysis of discretization methods for ordinary differential equations." Vol. 23. Berlin-Heidelberg-New York: Springer, 1973.

[56] R.L. Burden and J.D. Faires, "Numerical Analysis." Cengage Learning, 9th edition, 2010.

[57] S. Menzel, U. Böttger, M. Wimmer and M. Salinga, "Physics of the switching kinetics in resistive memories." Advanced functional materials (2015), 25(40), 6306-6325.

[58] F.A. Kröger and H.J. Vink, "Relations between the concentrations of imperfections in crystalline solids." Solid state physics. Vol. 3. Academic Press, 1956. 307-435.

[59] M. Schie, S. Menzel, J. Robertson, R. Waser, and R. De Souza, "Field-enhanced route to generating anti-Frenkel pairs in HfO_2." Physical review materials, 2(3), 035002 (2018).

[60] F. Stellari, E.Y. Wu, L. Ocola, T. Ando and P. Song, "Mapping and statistical analysis of filaments locations in amorphous HfO_2 ReRAM cells." Microelectronics Reliability (2023), 146, 114982.

[61] C.M.M. Rosário et al., "Metallic filamentary conduction in valence change-based resistive switching devices: the case of TaO_x thin film with x$\sim$1." Nanoscale 11.36 (2019): 16978-16990.

[62] C. Funck and S. Menzel, "Comprehensive model of electron conduction in oxide-based memristive devices." ACS Applied Electronic Materials 3.9 (2021): 3674-3692.

[63] C. Linderälv, A. Lindman and P. Erhart, "A unifying perspective on oxygen vacancies in wide band gap oxides." The Journal of Physical Chemistry Letters 9.1 (2018): 222-228.

[64] M.V. Ivanov, T.V. Perevalov, V.S. Aliev, V.A. Gritsenko and V.V. Kaichev, "Electronic structure of δ-Ta_2O_5 with oxygen vacancy: ab initio calculations and comparison with experiment." Journal of Applied Physics (2011), 110(2).

[65] K. Bao, J. Meng, J.D. Poplawsky and M. Skowronski, "Electrical conductivity of TaO_x as function of composition and temperature." Journal of Non-Crystalline Solids (2023), 617, 122495.

[66] A. Jain et al., "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation." APL materials 1.1 (2013).

[67] K.K. Kelley, "The Specific Heats at Low Temperatures of Tantalum Oxide and Tantalum Carbide1." Journal of the American Chemical Society 62.4 (1940): 818-819.

[68] K.T. Jacob, C. Shekhar and Y. Waseda, "An update on the thermodynamics of Ta2O5." The Journal of Chemical Thermodynamics (2009), 41(6), 748-753.

[69] C.L.T. Beechem, "Thermal Transport in TaOx Films for Memristive Applications." https://www.osti.gov/servlets/purl/1260377/, 2015.

[70] C. La Torre, "Physics-based compact modeling of valence-change-based resistive switching devices." Doctoral Dissertation, Rheinisch- Westfälische Technische Hochschule Aachen, 2019.

[71] T. Heisig et al., "Chemical Structure of Conductive Filaments in Tantalum Oxide Memristive Devices and Its Implications for the Formation Mechanism." Advanced electronic materials 8.8 (2022): 2100936.

[72] J. Woo et al., "Role of local chemical potential of Cu on data retention properties of Cu-based conductive-bridge RAM." IEEE Electron Device Letters 37.2 (2015): 173-175.

[73] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh and E. Eleftheriou, "Memory devices and applications for in-memory computing." Nature nanotechnology (2020), 15(7), 529-544.

[74] I. Boybat et al., "Neuromorphic computing with multi-memristive synapses." Nature communications 9.1 (2018): 2514.

[75] D. Ielmini and H.S.P. Wong, "In-memory computing with resistive switching devices." Nature electronics 1.6 (2018): 333-343.

[76] M. Abedin et al., "Material to system-level benchmarking of CMOS-integrated RRAM with ultra-fast switching for low power on-chip learning." Scientific Reports 13.1 (2023): 14963.

[77] C. Wang et al., "Ultrafast RESET analysis of HfO_x-based RRAM by sub-nanosecond pulses." Advanced Electronic Materials 3.12 (2017): 1700263.

List of publications

Peer reviewed journals

D.F. Falcone, S. Menzel, T. Stecconi, **M. Galetta**, A. La Porta, B.J. Offrein and V. Bragaglia, "Analytical modelling of the transport in analog filamentary conductive-metal-oxide/HfO_x ReRAM devices." Nanoscale Horizons (2024), 9(5), 775-784.

Conference proceedings

M. Galetta, D.F. Falcone, S. Menzel, A. La Porta, T. Stecconi, W. Choi, B.J. Offrein and V. Bragaglia, "Compact Model of Conductive-Metal-Oxide/HfO_x Analog Filamentary ReRAM Devices." ESSERC 2024-IEEE 50th European Solid-State Electronics Research Conference (ESSERC). IEEE, 2024.[2]

[2] *Just accepted*