DEPARTMENT OF ELECTRONICS AND
TELECOMMUNICATIONS ENGINEERING

MSc. in Nanotechnologies For ICTs

Master Degree Thesis

A Physical Model of Filamentary TaO_x/HfO_2 ReRAM Devices

Internal supervisor
Prof. Carlo Ricciardi

Candidate
Ludovico Carraria Martinotti

Local supervisors
IBM Research Zurich
Dr. Antonio La Porta
Research Staff Member in Neuromorphic Devices and Systems Group
Dr. Valeria Bragaglia
Research Staff Member in Neuromorphic Devices and Systems Group

Academic Year 2021-2022

Abstract

The advancement of information technology has increased exponentially the amount of data that is gathered and that requires processing, and the latter is predicted to become one of the most energy demanding operations in the very near future. Current data processing relies on Complementary Metal-Oxide-Semiconductor (CMOS) technology, but even the most energy efficient implementations suffer from an intrinsic power and speed inefficiency known as the von Neumann bottleneck, caused by the data transfer between memory and processing unit. It is therefore necessary to transition to new computing architectures and paradigms such as bio-inspired computing. The key component of these new solutions is the memristor, an element with multiple resistance levels that can be modified in a non-volatile manner. One of the most promising memristive technologies is filamentary valence-change mechanism (VCM) Resistive Random Access Memory (ReRAM), due to its scalability, ease of integration with current CMOS technologies and resistance range. The working principle of these devices is the creation of a conductive filament through an oxide layer sandwiched between two electrodes, followed by the modulation of the filament size to change the resistance. The main challenges in the development of ReRAM technologies are symmetry, linearity and stochasticity. A possible solution to these problems is the addition of a Conductive Metal Oxide (CMO) layer in the stack, though the exact role played by the CMO in the filament forming and the resistance switching is not yet fully understood.

This master thesis work is focused on the modeling of a CMOS-compatible bilayer ReRAM composed of substochiometric tantalum oxide (TaO_x) and hafnium oxide (HfO_2) sandwiched between two titanium nitride (TiN) electrodes. DC characterization, TEM analysis and CTLM measurements have been previously performed on devices with different TaO_x thickness and/or stochiometry and are used in this work as support for building the model. The selected simulation tool for this task is Ginestra®, a Kinetic Monte Carlo simulator developed by Applied Materials capable of emulating the behavior of multiple kinds of electronic devices. It allows to study how each parameter of the layers in the device stack affects the behaviour of the cell and extract the parameters from experimental data.

Capacitance-voltage measurements were performed to identify the metallic nature of the substoichiometric TaO_x and to build an effective model capable of capturing the effect of the CMO on both the forming and the set simulations.

Different modeling approaches were needed to obtain a structure that could match the ex-

perimental data. The analysis of the effect of different parameters on pristine state devices allowed to identify possible improvements of the fabrication process that could be implemented to fabricate devices with better performance.

The oxidation and reduction interplay between the metallic filament and the effective interface in this proposed model is demonstrated to be the phenomenon that governs the behavior of the devices by using bond breakage of the metal-oxide molecules, positive feedback Joule heating and recombination of oxygen ions and oxygen vacancies.

Acknowledgments

Before discussing the work proposed in this master thesis, I would like to thank all the people that contributed to this amazing experience and express my gratitude and appreciation for their support.

I would like to start by thanking all the members of the Neuromorphic Devices and Systems group at IBM Research Zurich for creating the amazing environment in which I had the chance to work for the past six months. In particular, I am deeply grateful to my supervisors Dr. Antonio La Porta and Dr. Valeria Bragaglia for their efforts in scientific discussions and feedback reviews. I would also like to give special thanks to Donato Francesco Falcone and Tommaso Stecconi for the great moments we shared, from scientific discussions to drinks at the bar. It was an incredible journey from start to finish and I am sure it would not have been the same in another group.

I would like to thank Prof. Carlo Ricciardi, Prof. Liliana Buda-Prejbeanu and Prof. Youla Morfouli for the management and supervision of the Master in Nanotechnologies for ICTs. This Master Degree has given me the opportunity to explore new environments, meet wonderful people and grow as a person, and for that I am thankful.

I am heartily thankful to my friends Niccolò, George and Simon, for supporting me during this journey and always being very close to me despite the travelling. Knowing that I can always count on you is a privilege I am very grateful for.

Vorrei ringraziare mia madre Silvia, mio padre Franco e mio fratello Giovanni. Non esprimo mai quanto il vostro affetto e il vostro sostegno siano importanti, vi sono grato per tutti i sacrifici che avete fatto per me e per tutto ciò che mi avete permesso di raggiungere grazie al vostro aiuto.

Concludo ringraziando mia nonna Secondina e mio zio Nelson, nonostante le avversità non avete mai smesso di fare il tifo per me e ve ne sono infinitamente grato.

The author acknowledges the Binnig and Rohrer Nanotechnology Center (BRNC) at IBM Research Europe - Zürich. The realization and characterization of the devices in this work were done within the framework of the funded European Union H2020 project “MANIC” (grant ID: 861153).

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Motivation

This chapter is dedicated to the explanation of the motifs behind the proposed work. Starting from the types and application of memristive devices, a description of Resistive RAM key mechanisms is presented. Then the goal of this modeling project is given, concluding with a summary of the various structural and electrical characterizations of the fabricated devices on which the modeling is based.

1.1 Memristive devices for new computing architectures

Current computing systems are based on complementary metal-oxide-semiconductor (CMOS) technology. The architecture of the devices, named after its creator John von Neumann in 1945, consists of having two separated units, one dedicated to computations and the other to memory. This separation of the core units has intrinsic power and speed problems related to the transfer of data between the two systems, and it is referred to as the von Neumann bottleneck. With the development of deep technology nodes to obtain smaller devices and the increase of data that requires processing, the von Neumann architecture is reaching its limits in multiple aspects: the reliability of the memory, due to the loss of non-volatility when structures are too small; the number of transistors on a single chip, as scaling of devices is reaching its lower bound; the heating of the chips, related to the number of transistors performing operations at an increasing rate; the power required to perform the computations, as more and more complex tasks are performed. The transition to brain-inspired architectures with a higher efficiency is therefore necessary to overcome these problems, especially for low power applications and processing of big-data for AI ap-

plications.

The key element of bio-inspired architectures is the memristor (memory resistor), a device that stores information in its resistance. The concept of the memristor, proposed for the first time by Chua in 1971 [1], differs from state of the art memory devices where information is tied to the accumulation of charges either in a floating gate for flash structures or in a capacitor for SRAM and DRAM.

The lowest and highest resistances that a memristor can have are referred to as low resistance state (LRS) and high resistance state (HRS), and by applying electrical stimuli it is possible to change the resistance of the device from one value to the other. A SET transition is the process of decreasing the resistance, while a RESET transition corresponds to its increase (see fig. 1.1).

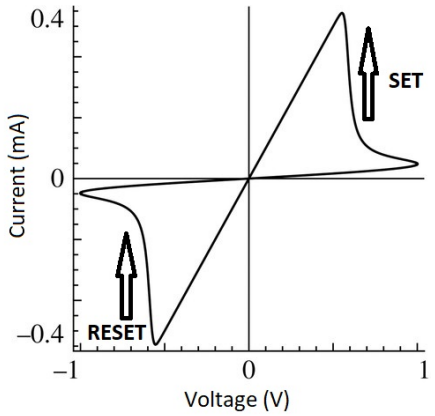


Figure 1.1: Example of standard memristive devices I-V curve, adapted from [1].

The main performance parameters of memristors are:

- operating voltages, as integration with CMOS technologies requires very low applied potentials;
- occupied area, as smaller devices allow to have more active components for the same chip size;
- data retention, related to the stability of the resistance level over time;
- endurance, defined as the maximum number of transitions from one state to the other until device failure;
- dynamic range (or On/Off ratio), defined as the ratio between HRS and LRS.

There are four types of memristive technologies that have received the most attention and that are most promising:

1) Phase change memristors (see fig. 1.2(a)), made of a crystalline insulator placed between a metallic heater and a metal electrode, in which the resistance change is tied to the modulation of the phase of the material between crystal and amorphous through Joule heating. The material in crystalline state has a low resistance, and by forcing a high current density with the metallic heater it is possible to reach a high temperature above the melting point of the insulator, which causes the phase to change to amorphous and obtain a RESET transition. To re-crystallize the material, by forcing a smaller current the temperature of the amorphous phase can reach the crystallization temperature to return to the LRS state and obtain a SET transition. The main problem of this technology is the operating voltage, since the RESET requires strong currents.

2) Ferroelectric Tunnel Junction (FTJ) memristors (see fig. 1.2(b)), made by a ferroelectric insulator sandwiched between two electrodes, in which the resistance modulation is tied to the alignment of the ferroelectric domains. By modulating the alignment of these domains through a voltage stimuli, due to the giant tunnel electroresistance effect the overall resistance of the devices is affected. The main problems of this technology are device area, to obtain a greater amount of domains, and dynamic range, as more domains translate in a larger resistance window.

3) Spin transfer torque (STT) memristors (see fig. 1.2(c)), made by a fixed and a free ferromagnetic layers separated by an oxide layer, in which the resistance modulation is tied to the alignment of the magnetization of the two layers. The modulation of the alignment of the free layer to the fixed layer can be achieved by applying a current, transfering the magnetization from one layer to the other. By giant magnetoresistance effect, when the two magnetizations are parallel the devices are in LRS, while it is in HRS when they are antiparallel. The main problem of this technology is the dynamic range.

4) Resistive switch memristors (see fig. 1.2(d)), typically made of an oxide layer placed between two metal electrodes. The switching of these devices is tied to the modulation of a conductive filament composed either of oxygen vacancies for OxRAM devices (also called

Valence Change Mechanism (VCM) cells), or of metal cations migrated from one electrode in the oxide for CBRAM devices (also called Electrochemical Mechanism (ECM) cells). The ion migration in the dielectric is governed by oxidation and reduction mechanisms. The main problem of this technology is due to the stochasticity of the filament modulation.

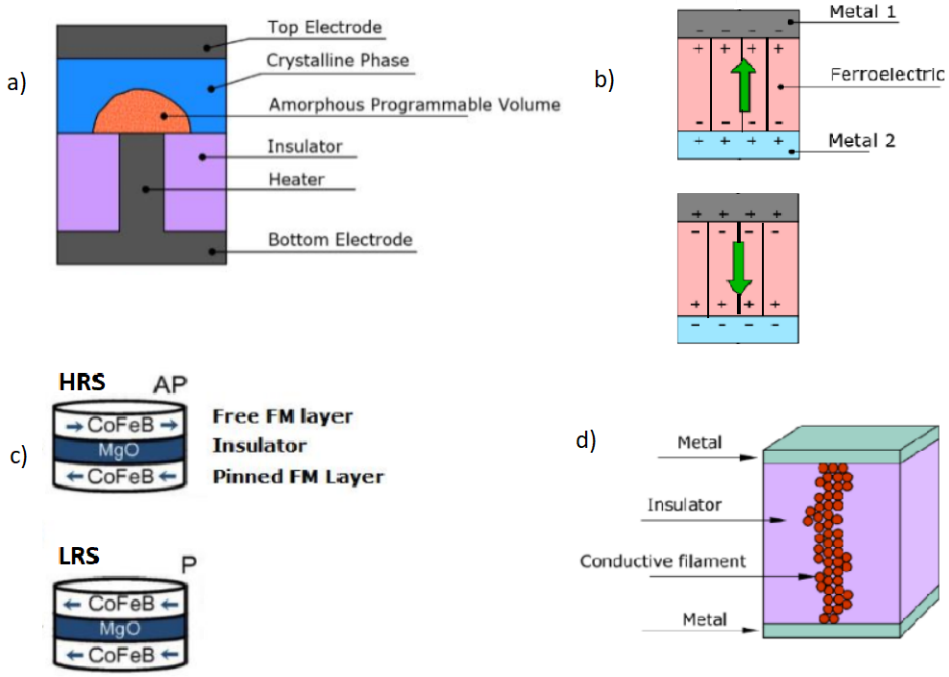


Figure 1.2: Schematic of the four most promising memristor types: phase change (a), ferroelectric tunnel junction (b), spin-transfer torque (c), resistive switch (d), adapted from [2].

To exploit the capabilities of memristors, a crossbar array (CBA) can be utilized. A schematic representation of this structure is shown in fig. 1.3. The co-location of memory and computing in a dense configuration allows to surpass von Neumann architectures. CBA structures are capable of performing very efficient vector-matrix multiplications, where the input voltages are the vector, the matrix is represented by the resistances of the memristors and the output vector is the output currents. This cannot be achieved in normal CMOS technology, where partial results have to be temporarily stored in dedicated memories to be transferred later again to the processing unit, a key advantage of this novel implementation.

Another great feature of CBAs is the possibility of co-integration with CMOS, as the fabrication processes are fully compatible with CMOS Backend-Of-Line (BEOL) fabrication. The most exciting implementation of CBAs and the context of this master thesis work is neuromorphic computing. A memristive array is capable of emulating the interconnection between two fully connected layers of neurons used in Artificial Neural Networks (ANN). Each memristor represents a synapse, while each input and output would be connected to an input or output neuron. The modulation of the resistance and the possibility to have multiple, stable levels corresponds to modulating the connection between input and output neuron, thus giving the ability to strengthen or weaken a neural connection. The result is a processing of information similar to the brain, where a dense and modifiable net of paths receives external stimuli and produces a response. The training of the neural network corresponds to modulating the resistance of the devices.

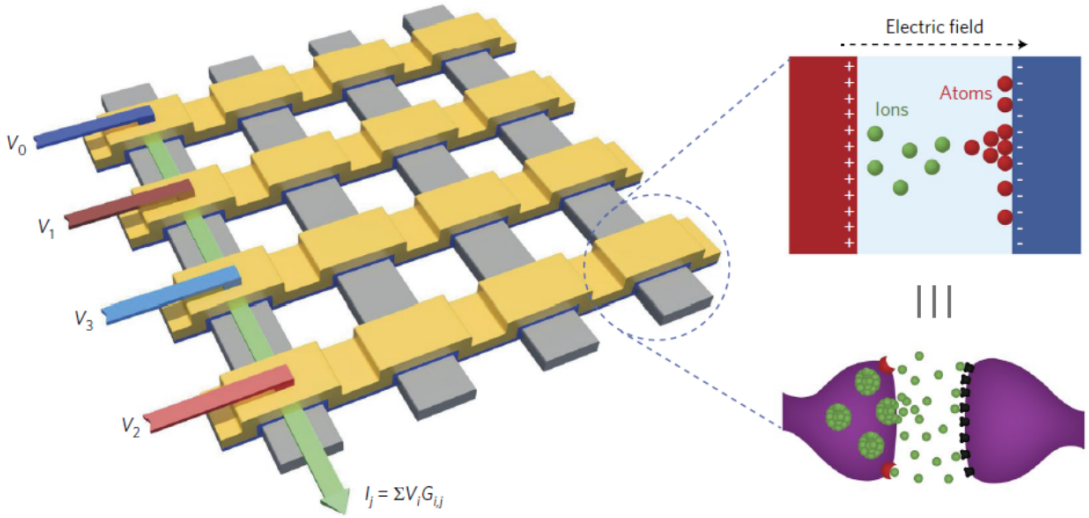


Figure 1.3: Schematic representation of a crossbar array structure [3].

Building hardware ANNs still remains a tough task, as many challenges have to be faced in order to increase the performance of memristors. Four main properties are required for this application:

- I-V symmetry, the symmetrical resistance change when applying positive or negative voltages to the device;

- Linearity, a linear stimuli-resistance change correlation. This is essential for the accuracy of the training, because the stochasticity of the resistance modulation causes lower performances of the network [12] or history dependant weight update [13];
- Stochasticity, the relative fluctuation of the resistance over the whole range of fluctuation. Low stochasticity equals a more accurate ANN [12];
- Multi-level operation, fundamental to resemble the plasticity of synaptic connections, though the actual number of levels is application dependent [13].

VCM memristors are a very promising candidate to match these requirements, though challenges related to symmetry, linearity and expecially stochasticity still have to be faced. The introduction of a conductive metal oxide (CMO) like TaO_x in the stack could be a potential solution to all of these problems thanks to its role in the operations of the devices, which is better explained in section 1.2. Building a model capable of emulating the behavior of these devices could prove essential in understanding the role of the conductive metal oxide, identifying the limiting factors of the technology and devising a way to overcome them by improving the fabrication of the devices, the training of the neural network or the overall structure of the implementation. It is important to first define in detail the concepts and mechanisms that dictate how ReRAM devices work, and this is explained in section 1.2.

1.2 VCM ReRAM operations

Once fabricated, VCM cells have a Metal-Insulator-Metal (MIM) structure, where the insulator is typically a large band gap poly-crystalline oxide. Devices in this state are referred to as being in pristine state. Due to the presence of an oxide layer, the resistance of the cells is higher than the one of the cells in their operating regime.
The first operation that the devices undergo is commonly referred to as forming, an oxide breakdown process by either forcing a large enough current (current forming) or applying a large enough voltage (voltage forming). The value of these electrical stimuli is strongly tied to the materials of each layer [14]. The breakdown of the oxide consists in the formation of a conductive filament that connects the two electrodes. The filament in VCM cells is made of oxygen vacancies, as the electric field in the oxide causes the metal-oxygen bonds to break. The oxygen ions then drift due to the electric field and accumulate to-

wards the electrode at the highest potential. It is commonly accepted that the beginning of the forming process happens along the grain boundaries of the oxide [15], because this region is rich of pre-existing structural defects and reactive sites. making it weaker than the structured region inside the grains. Once a high enough concentration of defects starts forming an efficient conducting path, the current density accumulates in this region, causing an increase of temperature due to Joule heating. The temperature increase weakens the metal-oxygen bonds due to thermal excitation, enhancing the rupture of the bonds by the electric field. This chain reaction leads to a transition from insulator to metal (see fig. 1.4), paired with the formation of a thick conductive filament, a current spike (see fig. 1.5) and a temperature spike in the region of the forming. This process is disruptive and irreversible, and without any quenching mechanisms the formed filament size would be too big to allow any resistance modulation. Therefore the forming process is performed with either a series resistance, to act as a voltage divider once the resistance of the device drastically decreases, or a series transistor (commonly known as 1T-1R) to get a current compliance and limit the positive feedback loop of the forming.

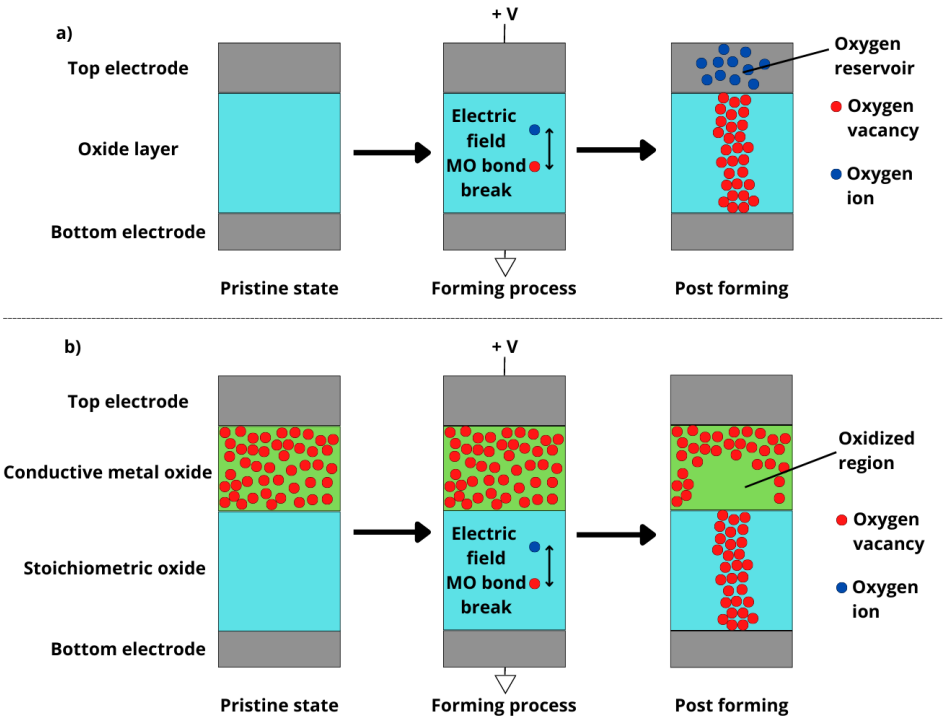


Figure 1.4: Schematic of the forming process for monolayer devices (a) and CMO bi-layers (b).

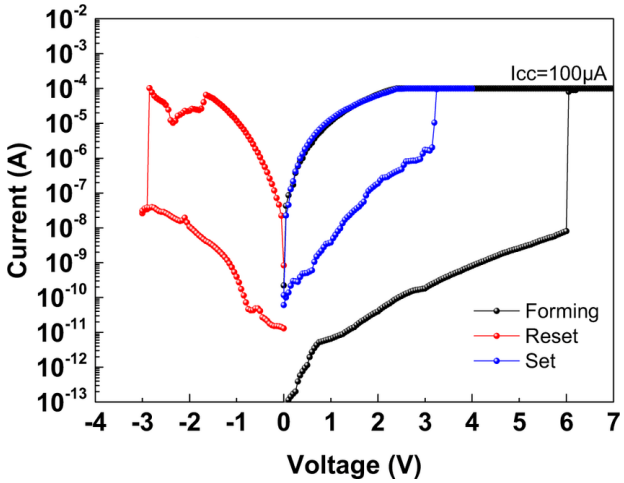
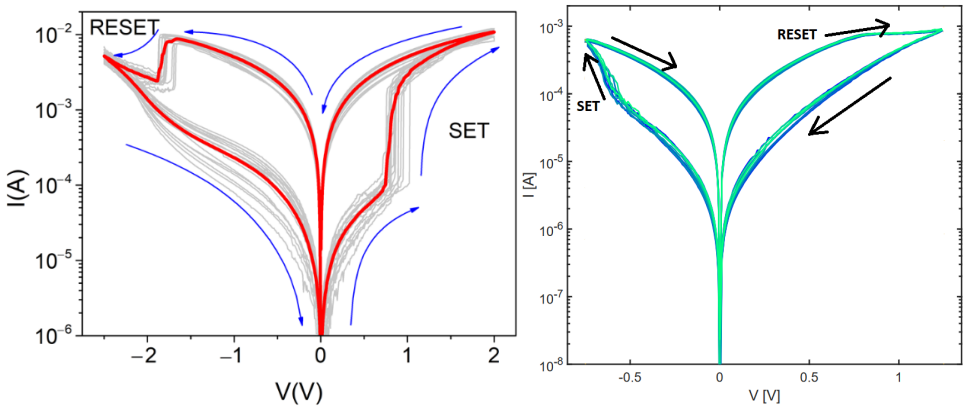


Figure 1.5: Example of forming, SET and RESET curve of an RRAM device with current compliance as a stopping mechanism, adapted from [4].

The migration of oxygen ions towards the highest potential causes a local oxidation above the filament. In standard RRAM cells, where only one stoichiometric oxide layer is present, the oxygen reservoir forms at one of the two electrodes. The idea behind the addition of a conductive metal oxide in the RRAM stack is to change the location of the oxidized area from the electrode to the conductive metal oxide, which would be the new oxygen reservoir, as shown in the comparison between fig. 1.4 (a) and (b). The novelty of these devices is the fact that the resulting state of the devices is HRS when forming with a positive feedback stop mechanism. This can be explained by the fact that there are two structures giving the overall resistance of the device, the filament and the oxidized cap, while in single oxide layer devices only the filament makes up for the resistance of the devices.

The SET and RESET operations are also different for the two types of devices. In single layer RRAM, by applying a negative voltage the oxygen ions of the reservoir are pushed to the bottom and recombine with the filament. The erosion of the filament causes a physical detachment from the top electrode, which results in a very steep current drop due to the newly formed gap, with the I-V curve shown in fig. 1.6 (a) and a schematic representation in fig. 1.7 (a) . The process is similar in bi-layers with the CMO, as the voltage drop on oxidized area causes the generation of oxygen vacancy/oxygen ion pairs and oxygen ions are pushed towards the bottom electrode, as shown in fig. 1.7 (b). The difference this time is that the transition tends to be more gradual, as the chemical reduction of the CMO counteracts the oxidation of the filament. The conductivity increase of the CMO outweighs

the conductivity reduction of the filament, thus giving rise to a gradual negative SET, as shown in fig. 1.6 (b).



(a) (b)

Figure 1.6: Example of SET and RESET I-V curves for monolayer RRAM devices (a) [5] and CMO bi-layer RRAM devices (b).

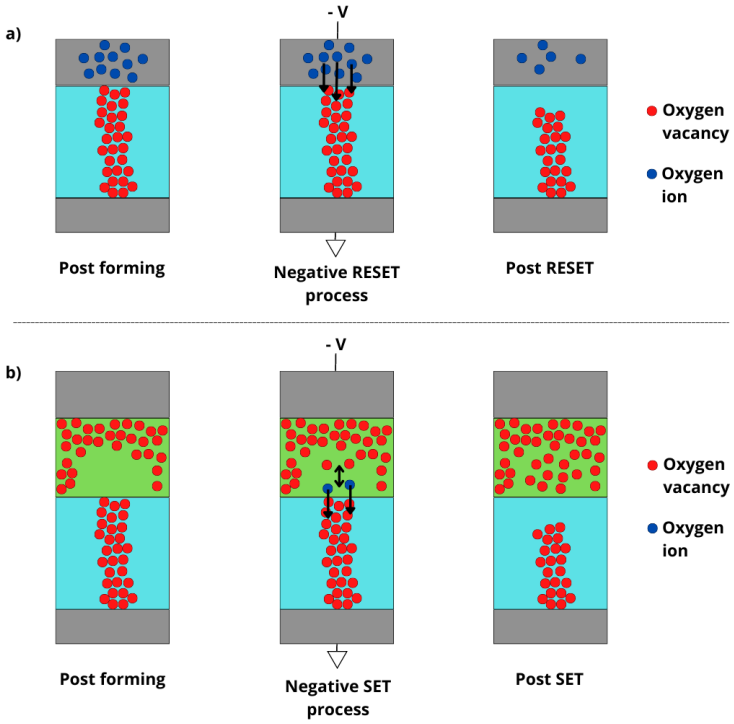


Figure 1.7: Schematic of the negative voltage RESET of monolayer devices (a) and negative voltage SET of CMO bi-layer devices (b).

After a negative voltage is applied, by applying a positive voltage in mono-layer devices the potential drops mostly in the most resistive region of the filament, which is the missing tip. This causes a transition in a process similar to the one of the forming, where oxygen vacancy/oxygen ion pairs are generated and the oxygen ions drift towards the oxygen reservoir in the top electrode. The filament is rebuilt and the conduction is strongly favored, thus giving a SET transition. A schematic of the process is shown in fig. 1.8 (a) and an example of I-V curves is shown in fig. 1.6 (a). In bi-layers with a CMO, the same process happens, except the oxygen returns in the CMO creating again an oxidized cap. The resistance increase of the CMO shadows the conductivity increase due to the formation of the filament tip, thus giving a gradual RESET transition. A schematic of the process is shown in fig. 1.8 (b) and an example of I-V curves is shown in fig. 1.6 (b). Due to the more gradual SET and RESET processes in bi-layers, the application of pulses instead of a DC voltages allows to obtain multiple intermediate states with lower stochasticity, better symmetry and linearity, which is the requirement for hardware AI implementations.

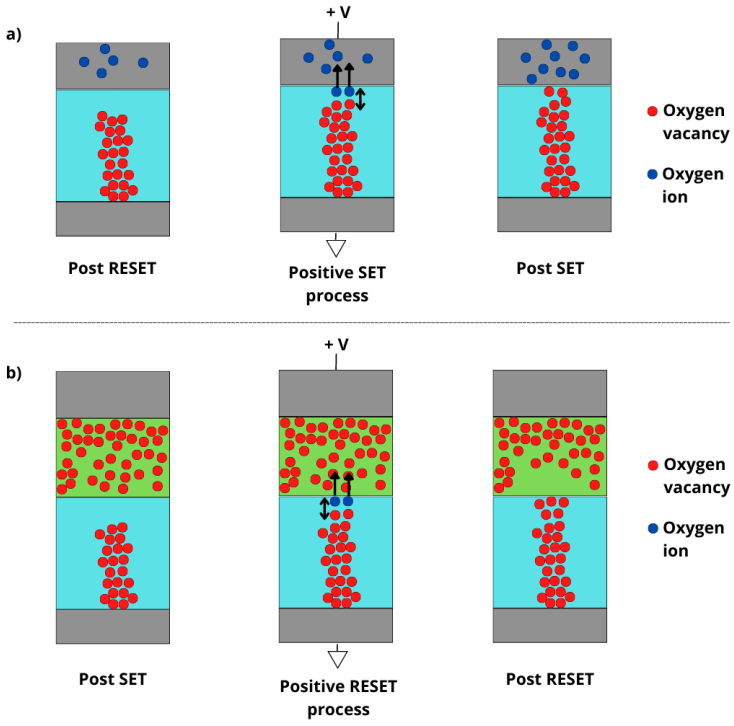


Figure 1.8: Schematic of the positive voltage SET of monolayer devices (a) and positive voltage RESET of CMO bi-layer devices (b).

While there are multiple models describing the mono-layer mechanisms of RRAM cells, from DFT calculations [16] to Kinetic Monte Carlo simulations [17] to compact models [18], the behavior of bi-layer cells with a conductive metal oxide has not yet been explored in detail, and the full potential of this new kind of technology has yet to be reached. The aim of the modeling would be understanding how to control the resistance modulation and improve the devices to match the implementation requirements.

1.3 Goal of the proposed work

The goal of this proposed work is the creation of a 3D full physics model capable of capturing the mechanisms of bi-layer RRAM devices with a CMO layer. The simulation tool chosen to emulate the devices is Ginestra®, an Applied Materials software [19]. The description of the pristine state, forming, set and reset of the devices is vital to understanding how to improve their performances. The aim is to identify possible material choices, process optimizations, operating conditions or stimuli application to have devices that match the requirements for the implementation of RRAM in crossbar arrays for ANNs.
In this master thesis work, an extensive study on TiN/HfO_2/TaO_x/TiN cells is carried out, covering the pristine state devices, the forming process and the initial negative set described in section 1.2. The novelty of the modeling is the presence of the substoichiometric tantalum oxide, and the main focus is on its effect on the behavior of the devices. Multiple characterization techniques have already been used to extract useful information on the devices, and their result is summarized in section 1.4.
This work will hopefully help in understanding the effect of the CMO in a bi-layer stack and give cues on how to create models of more complex systems.

1.4 Pre-existent device characterizations

Previous characterization techniques performed on TiN/HfO_2/TaO_x/TiN bi-layers with nominal HfO_2 thickness of 5.7 nm and nominal TaO_x thickness of 20 nm are summarized as follows:

- Transmission Electron Microscopy (TEM) analysis, a technique that allows to identify the thickness and the density of each layer in the stack. From fig. 1.9 it is clear that the

TaO_x layer is divided in two layers with different densities, 9.7 g/cm^{-3} for the interfacial layer and 10.7 g/cm^{-3} for the bulk TaO_x. This difference in density translates in different oxygen content as well as different resistivity of the two layers. The separation is due to the sputtering deposition technique, because at the beginning of the deposition the tool is not yet stable. The correlation between stoichiometry and density from [20] indicates a high oxygen deficiency, or in other terms a high concentration of oxygen vacancies.

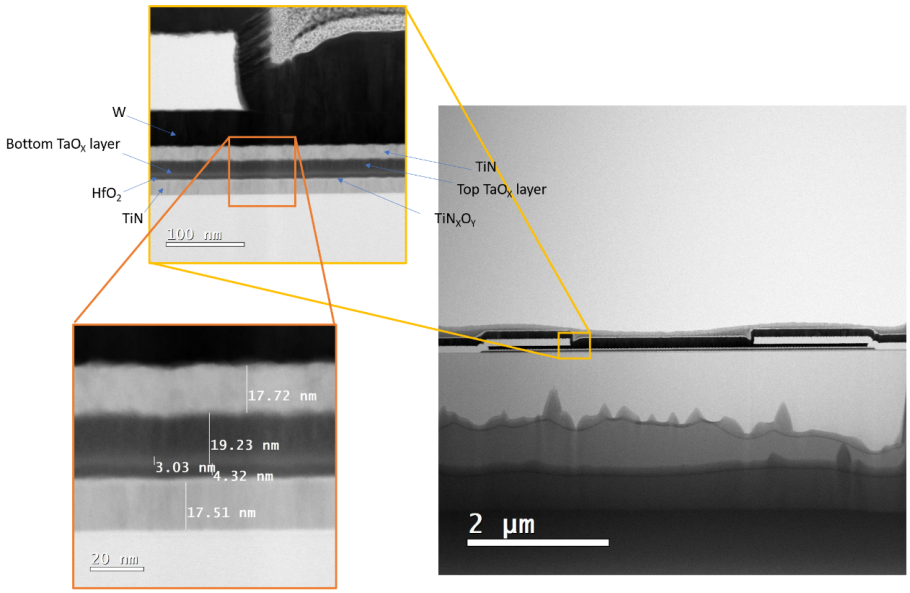


Figure 1.9: TEM characterization of the TiN/HfO_2/TaO_x/TiN RRAM devices, highlighting the split of the TaO_x in two layers with different densities [6][7].

- DC characterization of SET and RESET, showing in fig. 1.10 that for negative sweeps the devices switch from HRS to LRS and for positive sweeps the devices switch from LRS to HRS. Both transitions are gradual and by changing the stopping point of the sweep a different resistance can be obtained.

- Circular Transfer Length Method (CTLM) measurements of substoichiometric TaO_x deposited on a HfO_2/TiN/Si substrate with different sputtering chamber pressures allowed to extract both the sheet resistance and the resistivity of the bulk tantalum oxide layer (see fig. 1.11). The TaO_x sheet resistance of the devices used to fit the model is 30 kΩ/□.

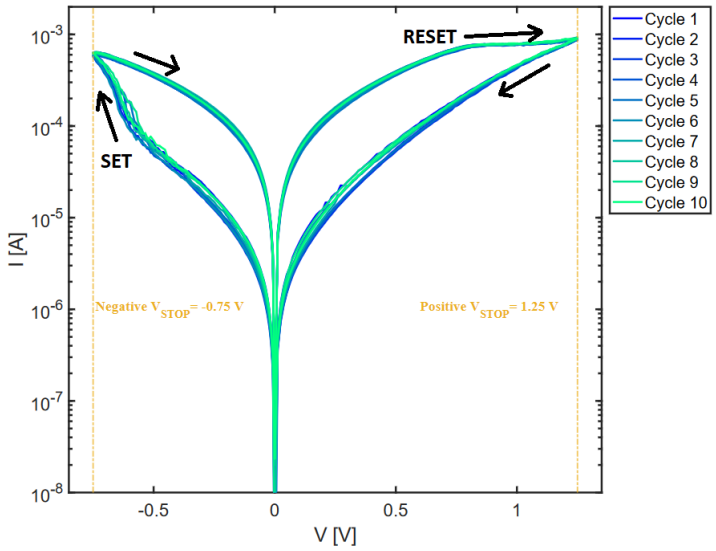


Figure 1.10: Result of electrical DC characterization, highlighting the negative SET and positive RESET of the devices [7].

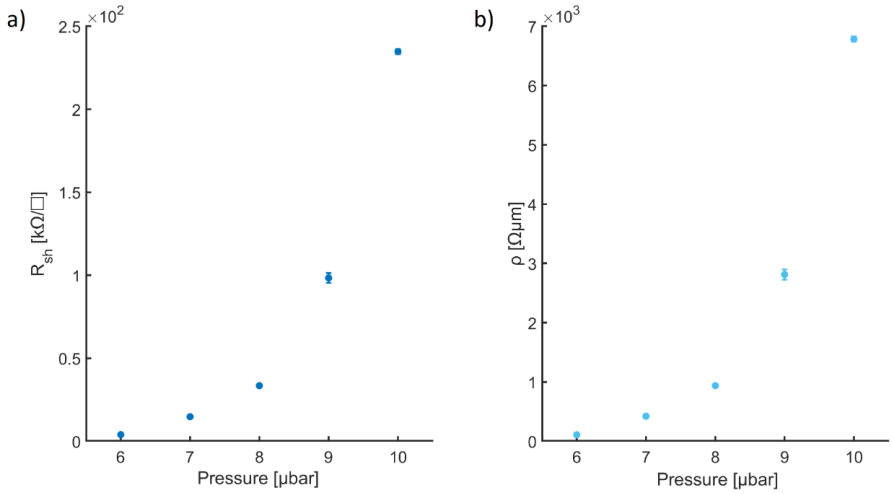


Figure 1.11: Data extracted from CTLM measurements to identify the trend of sheet resistance (a) and resistivity (b) for deposition chamber pressure [6].

Chapter 2

Simulation methods: Ginestra®

This chapter is dedicated to the description of Ginestra®, an Applied Materials software designed to simulate the operation and electrical characteristics of modern logic and memory devices such as FinFET, PCM and Ferro [19]. The tool is a 3D full-physics Kinetic Monte Carlo simulator designed to describe the steady state behavior and time evolution of the processes in RRAM cells with a multi-scale modeling approach. It allows to quantify all the mechanisms involved in conduction, specifically trap-assisted tunneling in filamentary RRAM devices, as well as material degradation and charge/ion transport to capture the operations that can be performed on the cells (forming, set, reset). This material is adjusted from the Ginestra® User Guide [8].

2.1 Device description

In this section a description of how devices are instantiated and simulated in the Ginestra® framework will be given, listing the device parameters and boundary conditions.
Ginestra® provides a built-in Metal-Insulator-Metal template that creates a single stoichiometric hafnium oxide (HfO_2) layer sandwiched by two titanium nitride (TiN) electrodes, as shown in fig. 2.1. It is possible to add layers, modify their thickness and change their physical characteristics to match the dimensions of the fabricated devices.

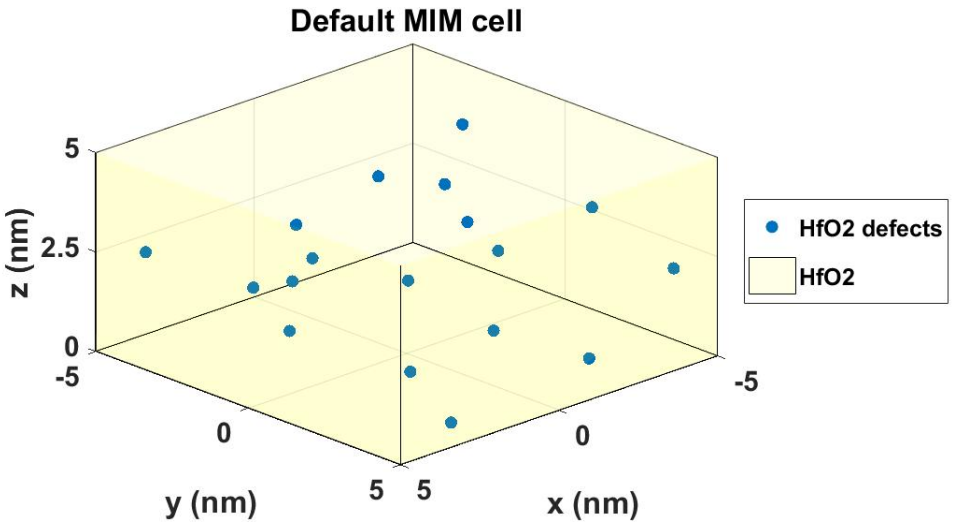


Figure 2.1: Ginestra® default Metal-Insulator-Metal cell.

2.1.1 Electrodes

To model electrodes, Ginestra® gives the possibility to choose between metal, semiconductor, ideal contact and insulator, and each of these choices has its own properties and boundary conditions. In the case of TiN/TaO_x/HfO_2/TiN stacks and standard RRAM devices both top and bottom electrodes are metals. They are described by three parameters:

- Work function Φ [eV]
- Electron density of states ρ [$m^{-3}J^{-1}$]
- Electron effective mass m^*

In Ginestra® metals are considered as perfect electrical conductors and as perfect thermal conductors, which means they do not show any resistance and their temperature is constant. This approximation can be applied when comparing the electrical and thermal resistances of the metallic layers and of the oxide layers, where the latter two are multiple orders of magnitude greater. This means that in the structural representation of the devices the electrodes are never shown but are always present (see fig. 2.1 as an example).

2.1.2 Oxide layers

The description of oxide layers is more complex than the one of electrodes since it has to cover all the mechanisms of RRAM devices. The relevant parameters for the modeling are:

- Dielectric permittivity ϵ_r

- Thermal conductivity λ $[Wcm^{-1}K^{-1}]$
- Electron and hole effective mass, m_h^* and m_e^*
- Electron and hole tunneling effective mass m_h^{tun} and m_e^{tun}
- Electron affinity χ [eV]
- Band gap BG [eV]

To complement these more general parameters, it is possible to introduce oxygen vacancy distributions and oxygen ion distributions. The movement of these species is described using a hopping model over an energy barrier between two adjacent sites. The diffusion rate R_D keeps into account both the effects of temperature and electric field and is defined as:

$$R_D = \nu \exp\left[-\frac{E_{AD}(x,y,z) - \gamma F_{eff}(x,y,z)}{k_B T}\right] \tag{2.1}$$

where

- ν [Hz] is a frequency pre-factor related to the bond oscillation frequency
- $E_{AD}(x,y,z)$ [eV] is the diffusion activation energy made by the x, y and z components
- γ [eÅ] is the field acceleration factor
- $F_{eff}(x,y,z)$ is the local effective electric field

As a modeling simplification, oxygen vacancies are considered fixed in space and only oxygen ions are allowed to diffuse in the oxide layers.

In the Ginestra® framework, oxygen vacancies behave as traps capable of capturing elecrons from the valence band or other traps. Their role in the conduction is governed by the following parameters:

- Defect energy mean E_{th} [eV], defined as the average distance between the energy level of the traps and the conduction band of the oxide
- Defect energy spread ΔE_{th} [eV], defined as the energy level range of the traps
- Charge q [e], as defects can act as charged particles based on the bonding of the oxide molecules. A trap that captures an electron is considered with the default charge lowered by 1

These parameters are key in describing the electrical behavior of the modeled filamentary RRAM devices after forming. In fact, the metallic filament is represented as a region with a

very high density of oxygen vacancies, which allows very efficient trap-assisted conduction. A summary of the relevant energy parameters is shown in fig. 2.2

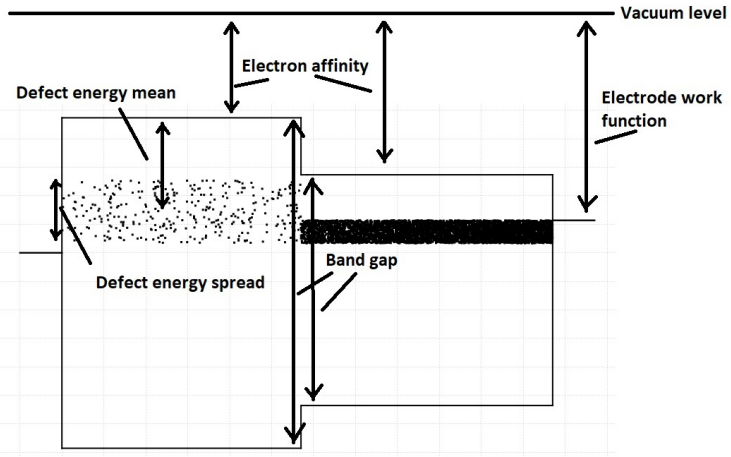


Figure 2.2: Example of a band diagram of a bilayer with the various energy parameters.

To describe the forming and switching mechanisms of RRAM cells, a generation/recombination mechanism involving the oxygen ion/oxygen vacancy pairs can be activated. The generation rate R_G and the recombination rate R_R are defined by exponential laws as:

$$R_G(x,y,z) = \nu \exp\left[-\frac{E_{AG} - \Delta E - bF(x,y,z)}{k_B T}\right] \tag{2.2}$$

$$R_R(x,y,z) = \nu \exp\left[-\frac{E_{AR}}{k_B T}\right] \tag{2.3}$$

where for generation

- ν [Hz] is the bond vibration frequency prefactor, that can differ from the one of equation (2.1)
- E_{AG} [eV] is the zero-field bond-breakage activation energy, defined as the energy necessary to activate the generation process in absence of an electric field
- $b = p_0(2 + \epsilon_r)/3$ [eÅ] is the bond polarization factor, where p_0 [eÅ] is the polarizability and ϵ_r is the dielectric constant of the oxide, and it represents the strength of the bond between metal and oxygen in the oxide
- ΔE [eV] is a reduction of activation energy based on the proximity of other defects

while for recombination

- E_{AR} [eV] is the recombination activation energy

Typically the activation energy of the recombination process E_{AR} is very low (below 0.5 eV) compared to the one of the generation process E_{AG}, which is tied to the material and its internal bonds. This is because oxygen vacancies are highly reactive sites that require an oxygen ion to reach a neutral state of lower energy, therefore this process is strongly favored, while the generation process requires the breaking of a metal-oxygen atomic bond either by the effect of the local electric field or the temperature.

2.1.3 Boundary conditions

Together with the boundary conditions imposed by the choice of electrode type, so fixed temperature and negligible series resistance of the metallic electrodes, other boundary conditions are applied in order to solve the physical equations.
The first is an electrical boundary condition on the potential that reads as

$$\Phi = V - \Phi_{MS} \tag{2.4}$$

where

- Φ [V] is the effective potential across the two electrodes
- V [V] is the applied voltage
- Φ_{MS} [V] is the difference of work function between the two electrodes

The tuning of the electrodes parameters is therefore important to understand how the voltage is distributed on each layer and across the device.

The second is a thermal boundary condition related to the sidewalls of the simulated devices and reads as

$$\nabla \Phi_{Th} = 0 \tag{2.5}$$

where Φ_{Th} is the thermal flux.
The divergence of the thermal flux is null, so the heat is confined in the device. The simulated device size could have an impact on the maximum temperature reached due to heat reflection at the edges of the cell. This is a critical aspect in forming simulations that could lead to incorrect results.

It is possible to let the diffusing species escape the cell by selecting leaky interfaces between electrodes and oxides, but in this proposed modeling this option was never explored.

2.2 Conduction mechanisms

In this section the conduction mechanisms considered by Ginestra® will be described, with specific focus on the relevant ones for the modeling of the HfO_2/TaO_x devices. A schematic summary is shown in fig. 2.3.

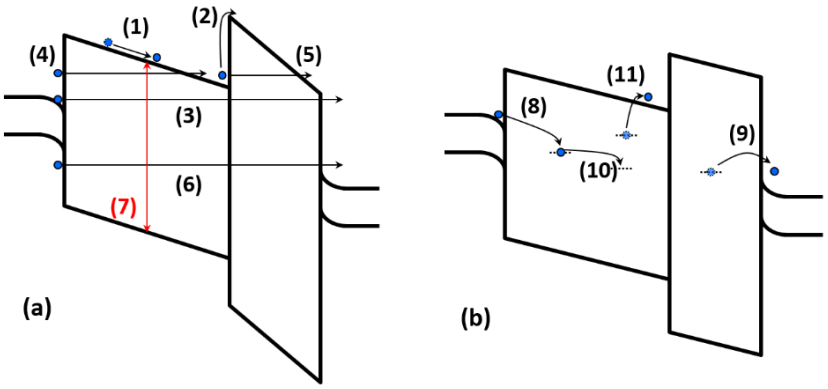


Figure 2.3: Schematic representation of the different electron transport mechanisms considered in Ginestra®: (a) tunneling and others; (b) trap-assisted. The mechanisms are numbered as follows: (1) drift-diffusion; (2) thermionic emission; (3) direct tunneling; (4) electrode-band tunneling; (5) intra-band tunneling; (6) band-to-band tunneling; (7) local generation-recombination; (8) electrode-to-trap capture; (9) trap-to-electrode emission; (10) trap-to-trap transition; (11) trap-to-band transition [8].

2.2.1 Drift/diffusion and tunneling mechanisms

In reference to fig. 2.3 (a), the electron drift-diffusion and tunneling conduction mechanisms considered are

- (1) Drift-diffusion, the movement in the conduction band of a single layer, negligible in the case of large band gap oxides used in RRAM technologies
- (2) Thermionic emission, the process of moving from the conduction band of an oxide layer to the conduction band of another one by thermal excitation, that can cause asymmetry in the I-V curves for positive and negative voltages

- (3) Direct tunneling from one electrode to another, a phenomenon that only matters when the thickness of the stack is of a few nanometers
- (4) Electrode-band tunneling, from one electrode to the conduction band of an oxide layer, which comes into play when the band bending causes an effective lowering of the tunneling barrier
- (5) Intra-band tunneling, from the conduction band of an oxide to the conduction band of another one, which again only happens for high enough band bending
- (6) Band-to-band tunneling, from the valence band of a layer to the conduction band of another
- (7) Local generation-recombination by thermal or photonic excitation

2.2.2 Trap-assisted mechanisms

In reference to fig. 2.3 (b), the electron trap-assisted conduction mechanisms considered are

- (8) Electrode-to-trap capture, from electrode to an unoccupied trap
- (9) Trap-to-electrode emission, from an occupied trap to an electrode
- (10) Trap-to-trap transition, from an occupied trap to an unoccupied one
- (11) Trap-to-band transition, from an occupied trap to the oxide conduction band by either thermal or photonic excitation

Except for (11), all of these mechanisms are the basis on which the model is built. For pristine devices, the transport is dominated by the very low concentration of defects in the oxides, as the wide band gaps do not allow for non-assisted tunneling and drift-diffusion. After forming, since the filament is a region with a critically high density of defects, trap-to-trap transitions allow for a very efficient conduction, like in a real metal. The key to the efficiency of the conduction is the energy alignment of the traps of the oxide layers and the Fermi level of the electrodes.

2.3 Simulation options

In this section the simulation options useful for the RRAM modeling offered by Ginestra® are explained.

2.3.1 Electrical options

In the electrical options section, it is possible to define the voltage applied to the device by specifying the shape of the input and the amplitude for DC simulations. In transient simulations, the time evolution of the signal has to be specified instead, and it is possible to add a stopping condition (like maximum device current) that stops the simulation once it has been reached. It is also possible to specify the initial temperature (usually fixed at 300 K) and external components connected to the device, like in the case of RRAM crossbar arrays either series resistances to limit the post-forming voltage or series transistors to limit the post-forming current.

2.3.2 Physical options

In the physical options section, it is possible to specify what physical equations Ginestra® has to solve in the test. Multiple subgroups of settings can be modified, and the relevant ones for RRAM simulations are

- Solution dimensionality, which allows to select along which axis to solve the Poisson and Fourier equations
- General, where it is possible to use simplified approaches for faster simulations like a defect density approach instead of considering discrete defects, or the use of Boltzmann statistics instead of Fermi-Dirac
- Electron-hole transport, where it is possible to select which conduction mechanisms to simulate
- Poisson's equation, where it is possible to enable the contribution of the charge of defects to the potential calculation
- Fourier's equation, with the option of using the steady-state approximation
- Degradation, where it is possible to enable defect generation/recombination and local modifications of the dielectric constant and the thermal conductivity based on the defect concentration. When the defect density reaches the critical value of 10^{22} cm^{-3}, both dielectric permittivity and thermal conductivity of the highly defective volume are set to metal-like values

2.3.3 Design of experiment

The design of experiment (DOE) is a very useful tool that allows to perform the same test on the same device while changing either one or a combination of parameters. DOEs can be used to explore how different device characteristics may impact on the overall device performance, like in the case of the variation of the thickness of an oxide layer, and to identify how multiple parameters interplay with each other, like in the case of band gap and oxygen vacancies mean energy level.

Chapter 3

Characterization methods

This chapter is dedicated to the experimental part of the master thesis. The aim of these measurement techniques is the extraction of useful parameters for the simulations.

3.1 Quasi-Static Capacitance-Voltage measurement

A Quasi-Static Capacitance-Voltage measurement (QS-CV) of a ReRAM device consists of the superposition of a quasi-static DC sweep and a sinusoidal AC signal. The measured quantity is the AC current flowing through the device, from which it is possible to extract the impedance and subsequently the capacitance of the Device Under Test (DUT). The measurement and cabling setup of the QS-CV measurement is shown in fig. 3.1. The multi-frequency capacitance measurement unit (MFCMU) is an accessory of the parameter analyzer $Agilent B1500A$ that allows to perform IV and CV characterization at the same time when used in the four terminal (4T) configuration [9]. The high current terminal Hc applies the superimposed DC and AC voltages, and the high potential terminal Hp senses the actual AC signal applied to the DUT. The low current terminal Lc sinks the DUT current through a reference resistor and a negative feedback loop in the parameter analyzer maintains the low potential terminal Lp as close as possible to a null potential, to create a virtual ground. The Hc and Hp terminals are connected to the probing manipulator through a coaxial cable, and the Lc and Lp terminals are connected to the substrate of the device through another coaxial cable and the sample holder of the probe station. The guard of the coaxial cables is connected to form a shielded 2T configuration, which allows to have more precise measurements since the residual inductance of the cables is known. The voltage is

always applied at the top electrode of the ReRAM devices (CH), while the bottom electrode is deposited on the conductive substrate, so it is possible to connect the low potential/low current terminal directly to the substrate (CL). The overall applied voltage is kept at low values compared to the forming voltage, to avoid any unwanted long term degradation of the oxide layers. The positive and negative DC sweeps consist both of 101 steps of 30 mV, each lasting for 10 ms, from 0 to 3 V and from 0 to -3 V respectively. Positive and negative sweeps are applied separately to avoid charge accumulation at the interface of each layer. The applied AC signal is a 50 mV sinusoidal stimuli at a 1 kHz frequency. The impedance Z_x is measured simply as $Z_x = V_x/I_x$, where V_x is the applied AC signal and I_x is the AC current sensed at the Lc terminal. The conductance is also measured thanks to the 4T MFCMU configutation, and this allows to extract the capacitance of the DUT.

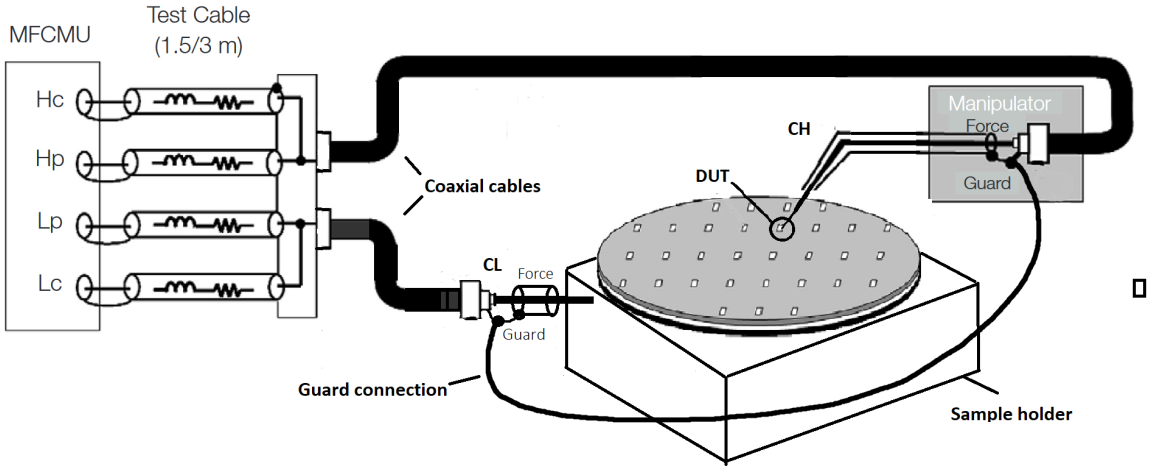


Figure 3.1: Schematic of the QS-CV measurement setup, adapted from [9].

3.2 Thermal conductivity measurement

The thermal conductivity extraction for thin films, also called 3Ω method, is a procedure that exploits the electrically and thermally insulating nature of the oxides in the ReRAM stack. A schematic representation of the measured structures is shown in fig. 3.2.

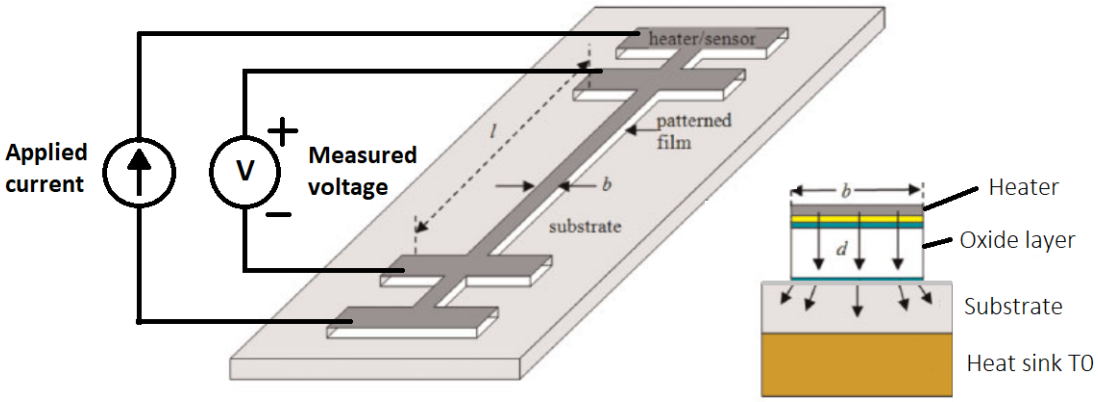


Figure 3.2: Schematic of the structures measured to extract the thermal conductivity of the oxide layers. The solid arrows of the cut section represent the heat flux through the stack.

All the electrical measurements performed are 4 probe measurements, hence the presence of four pads in the structure, 2 to apply the current and 2 to measure the voltage. It is important to have a structure with vertical walls to avoid lateral spread of the heat flux, and this is achieved by using the heater as a lithographic mask when etching the layers. In this configuration, the thermal resistance R_T of the oxide layer, given l and b the length and the width of the heater, d the oxide thickness and λ the oxide thermal conductivity, is defined as

$$R_T = \frac{d}{\lambda bl} \tag{3.1}$$

The two physical phenomena relevant for this method are:

- Joule heating of a heater, dictated by the equation

$$R(T_0 + \Delta T_h) - R(T_0) = \beta R(T_0) \Delta T_h \tag{3.2}$$

  where $R(T)$ [Ω] is the temperature dependant resistance of the heater, ΔT_h [K] is the heater temperature increase, T_0 [K] is the heat sink temperature and β [K^{-1}] is the thermal coefficient of the heater.

- Heat flux from heater to heat sink through the oxide layer (see fig. 3.2), dictated by the equation

$$\Delta T_h = (R_{Tsub} + R_{Ti} + R_T + R_{Th}) N_h \tag{3.3}$$

where ΔT_h [K] is the heater temperature increase, R_{Tsub} [W K^{-1}] is the thermal resistance of the substrate, R_{Ti} [W K^{-1}] is the thermal resistance of the multiple interfaces between each layer, R_T [W K^{-1}] is the thermal resistance of the layer under examination, R_{Th} [W K^{-1}] is the thermal resistance of the heater and N_h [W] is the power dissipated by the heather through the stack.

It is important to note that the thermal resistance R_T and the electrical resistance R(T) are different quantities and affect different physical effects.

The thermal coefficient of the heater can be extracted by measuring at different temperatures the resistance of the device with a low DC voltage, typically 0.2 V. For metals, the resistance increases linearly with temperature due to thermal expansion. A typical measurement curve is shown in fig. 3.3.

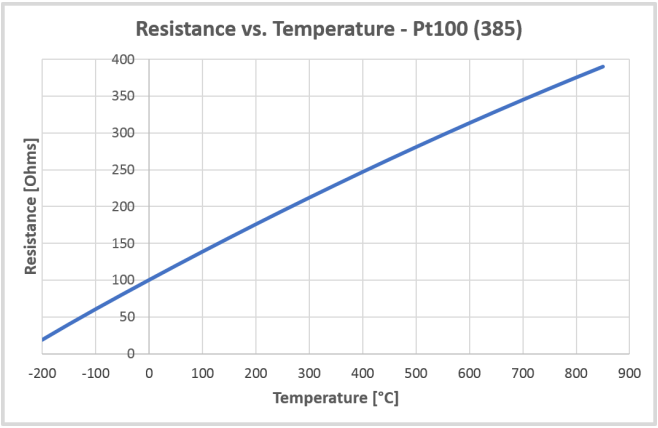


Figure 3.3: Example of an R-T measurement of a metallic heater [10].

By having structures with different oxide thickness, like in the case of fig. 3.4, it is possible to isolate the contribution of the oxide layer to the overall thermal resistance. Given d_1 and d_2 two different oxide thickness values, T_1 and T_2 the temperatures reached by the two respective heaters and N_1 and N_2 the power dissipated, these quantities are related by eq. 3.4, which can be extracted by combining eq. 3.1 and 3.3.

$$\frac{(T_2 - T_0)}{N_2} - \frac{(T_1 - T_0)}{N_1} = \frac{(d_2 - d_1)}{\lambda b l} \tag{3.4}$$

The dissipated power is calculated as the product of current and voltage, while the resistance and therefore the temperature of the heater can be calculated by their ratio. Since the structural parameters are known from fabrication processing, it is possible to extract the

thermal conductivity λ. The strength of this method is the fact that all the contributions of the other layers and of the interfaces to the thermal resistance present in eq. 3.3 are neglected, as they only act as an offset to the linear relation between oxide thickness and oxide thermal resistance. A typical result of the measurement procedure is shown in fig. 3.5, where the slope of the curve is proportional to the inverse of the thermal conductivity.

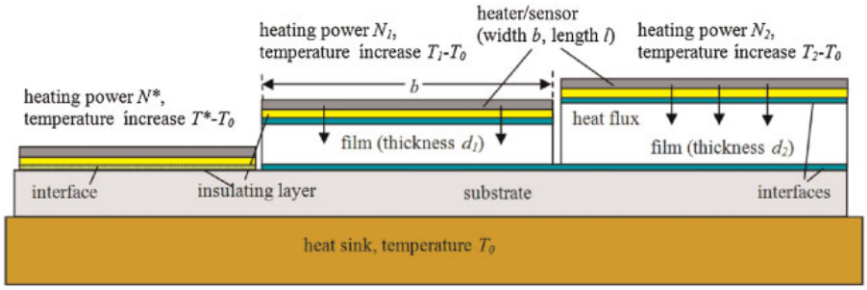


Figure 3.4: Schematic representation of stacks with different oxide thickness to have different oxide thermal resistances [11].

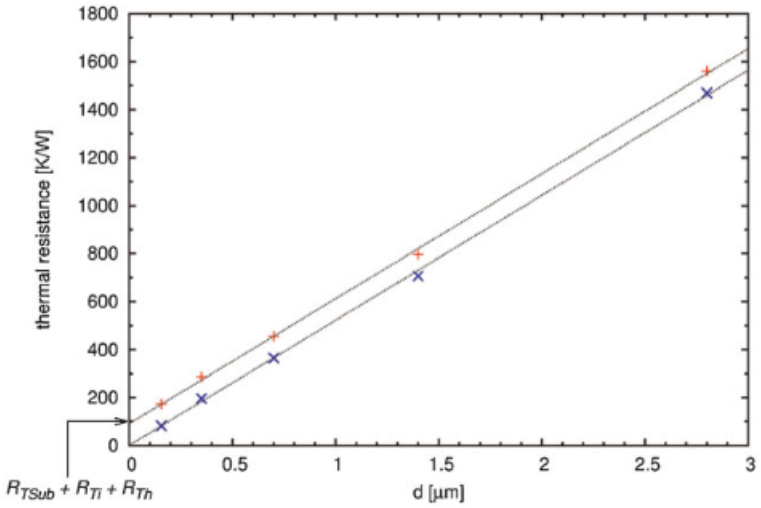


Figure 3.5: Example of the result of the thermal conductivity measurement with the 3Ω method [11].

Chapter 4

Results and discussion

This chapter is dedicated to the results obtained from measurements and simulations, with explanations to justify them. Starting from C-V measurements and previous results shown in section 1.4, it was possible to build a Kinetic Monte Carlo model of pristine state, forming and negative set of the hafnium oxide and tantalum oxide based RRAM devices.

4.1 C-V measurements

This section is dedicated to the results of the C-V measurements on the TiN/HfO_2/TaO_x/TiN devices. The goal of this experiment is to evaluate the dielectric constant of the oxide layers. To do so, three chips with different nominal thicknesses but same nominal CMO resistivity have been characterized with the measurement described in section 3.1. All the other layers of the stack are deposited using the same deposition conditions. This allows to identify the dielectric constant of the thick TaO_x by interpolation of the measured values with a series capacitance equation.

4.1.1 Device measurements

Ten devices with nominal TaO_x sheet resistance of 30 k$\Omega/\square$ have been measured using the setup shown in fig. 3.1. The measured devices are in pristine state. The measurements are shown in fig. 4.1.
The voltage dependence of the capacitance of the devices and the difference between positive and negative biases could be explained by the reduction of the effective thickness of the oxides [21]. The asymmetry could be explained by the presence of tantalum species in the

substoichiometric TaO_x layer. Since the work function of tantalum is 4.1 eV [22] and the work function of titanium nitride is between 4.3 and 4.7 eV, a built-in electric field could cause a larger capacitance change at negative voltages compared to the one at positive voltages.

The device to device variability is more pronounced for thinner TaO_x layers, as the process variation from one structure to the other and the surface roughness has a bigger impact in smaller structures.

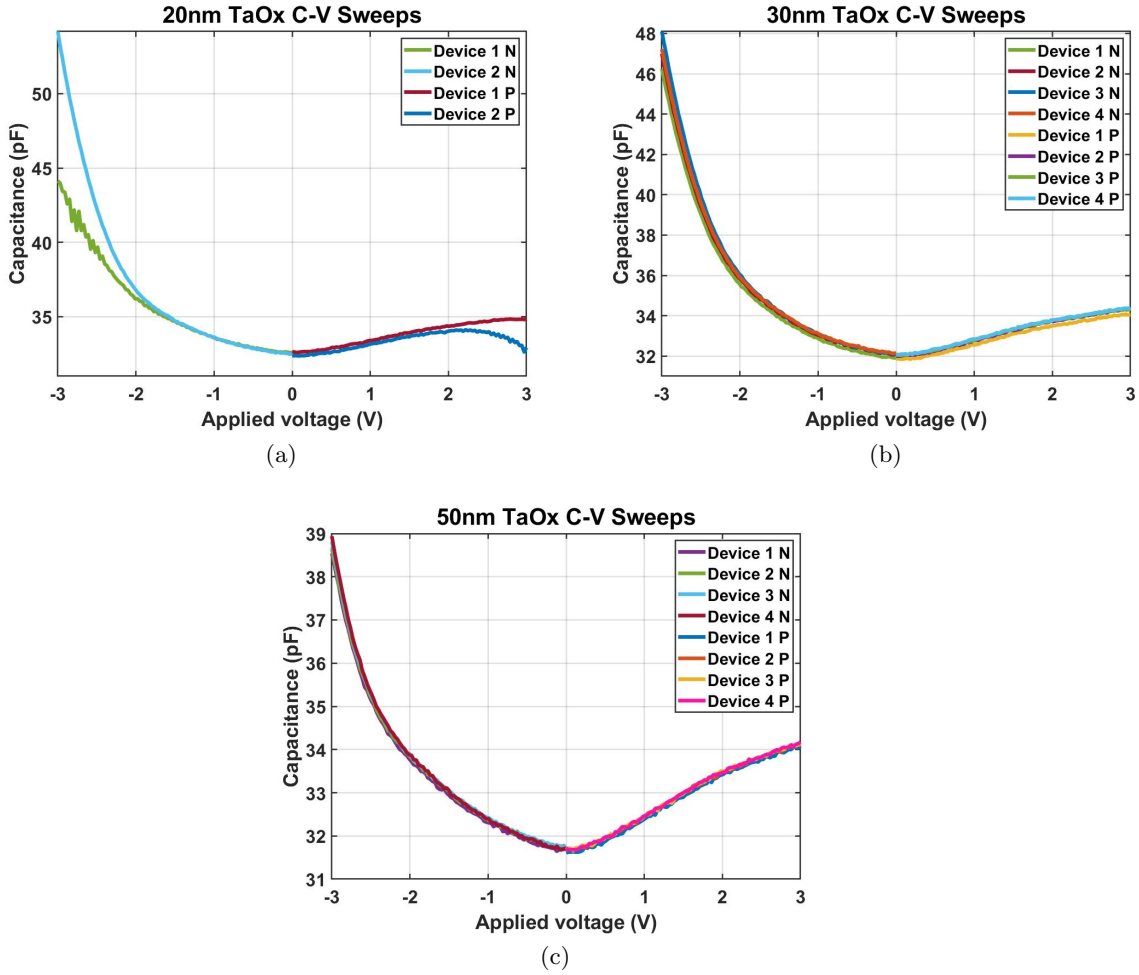


Figure 4.1: Capacitance-Voltage measurements of devices with different nominal TaO_x thickness for AC signal frequency 1 kHz and amplitude 50 mV. (a) 2 devices with 20 nm, (b) 4 devices with 30 nm and (c) 4 devices with 50 nm.

4.1.2 Zero-field dielectric constant extraction

From the measurements of section 4.1.1 it is possible to see the correlation between TaO_x thickness and normalized capacitance over the area for a null DC voltage. A point plot of the correlation is shown in fig. 4.2, where it is clear that a thinner oxide stack increases the capacitance. The relative difference for different oxide thickness is small compared to the absolute values. This can be explained by the conductive nature of the substoichiometric TaO_x, which entails that the capacitance value is dominated by the HfO_2 and the thin interfacial TaO_x.

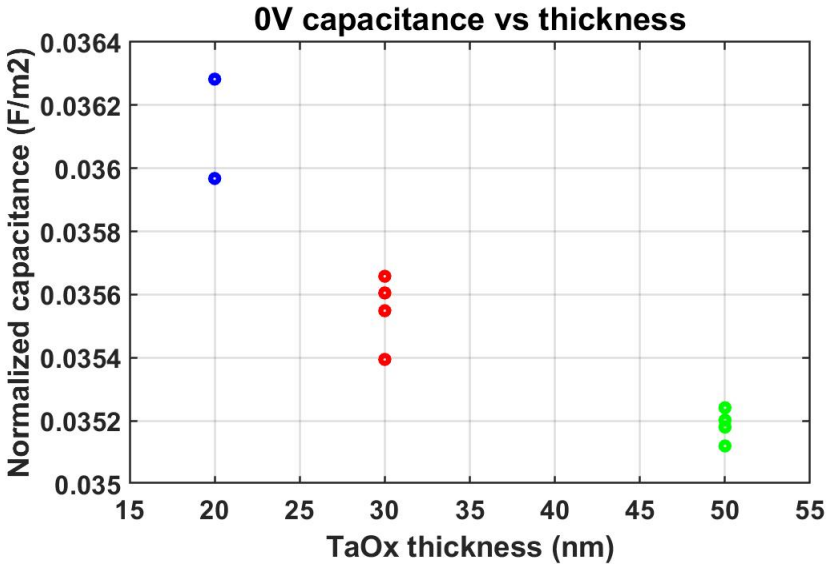


Figure 4.2: Normalized capacitance measurement at 0 V as a function of the nominal thickness of the TaO_x layer.

To extract the dielectric constant of the TaO_x layer, it is possible to use a simplified series capacitance equation. The normalized capacitance C_σ is extracted from the following equation:

$$C_\sigma = \frac{\epsilon_0}{\frac{t_{HfO_2}}{\epsilon_{r,HfO_2}} + \frac{t_{B-TaO_x}}{\epsilon_{r,B-TaO_x}} + \frac{t_{T-TaO_x}}{\epsilon_{r,T-TaO_x}}} \tag{4.1}$$

with t indicating the thickness of a layer, ϵ_r the dielectric constant of the layer and ϵ_0 the vacuum permittivity. For HfO_2 and the interface layer the thicknesses are estimated by the deposition times and the TEM imaging, and are considered 5.7 nm and 3 nm respectively. The dielectric constant of HfO_2 is considered as 22 and for the interface layer 28, a value obtained in [23] for HfTaO layers, as some intermixing could happen in the deposition

process. By interpolating the mean value for each bulk tantalum oxide thickness and using the values of the other two layers as constants, the extracted dielectric permittivity of TaO_x is around 5000. This value leads to the hypothesis that the actual behaviour of the TaO_x layer is more similar to the one of a metal. This is also in line with the resisitivity measurements, as well as the density and the stoichiometry of the layer [20].

4.2 Ginestra® simulations

This section is dedicated to the description of the models of the devices on Ginestra®. The goal is to obtain a holistic model capable of describing forming and switching of the device while matching the experimental data. The experimental data used as a comparison is the median measurement of devices with nominal TaO_x thickness of 20 nm and device area 30x30 μm^2 for both pristine state simulations and switching simulations.

4.2.1 Tri-layer simulations

The first model is based on the TEM analysis, resistivity measurements and DC characterization shown in section 1.4, C-V measurements of section 4.1 and results obtained in [20]. It is a three layer structure in which the low stoichiometry of the top TaO_x is modeled as a very high concentration of oxygen vacancies.

4.2.1.1 Device representation and parameters

A representation of the simulated device is shown in fig. 4.3, and a summary of the parameters is given in table 4.1. For the HfO_2 layer, the parameters are the default ones from the Ginestra® material library, and the defect density is in line with the one of other stoichiometric oxides analyzed in other studies [24]. The defect density of the interface layer is two orders of magnitude greater than the one of the stoichiometric HfO_2 layer because of the density measured by the TEM.
Different band gaps, electron affinities and defect densities are considered for interface and T-TaO_x to model the density and conductivity difference between the two layers. The choice of parameters is in line with literature values [25].
The area of the device is set to 10x10 nm^2 and the thickness of the T-TaOx layer is set to 10 nm to limit simulation times, as the software speed decreases exponentially with the number

of defects. The thickness of HfO_2 and interface is 5.7 nm and 3 nm respectively, as obtained from TEM analysis and as used in equation 4.1 to extract the dielectric permittivity. The top and bottom electrodes are default TiN of the Ginestra® material library.

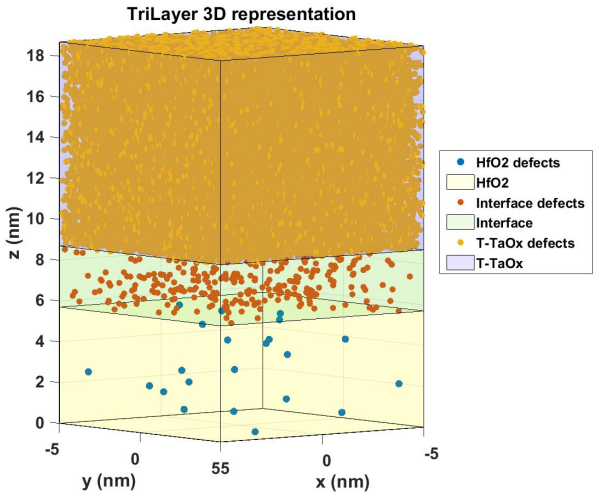


Figure 4.3: Tri-layer cell 3D representation.

	HfO_2	Interface	T-TaO_x	TE	BE
Work Function [eV]	-	-	-	4.8	4.57
Density of states [m^{-3}/J]	-	-	-	$9*10^{40}$	$9.037*10^{45}$
Band Gap [eV]	5.8	4.2	4.2	-	-
Electron affinity [eV]	2.5	3.2	3.2	-	-
Dielectric permittivity	22	50	5000	-	-
Thermal conductivity [$Wcm^{-1}K^{-1}$]	0.005	0.01	0.01	-	-
Tunneling mass [m_0]	0.25	0.3	0.3	-	-
Defect density [cm^{-3}]	$5*10^{19}$	$4*10^{21}$	10^{22}	-	-
Defect energy mean [eV]	2.05	1	1	-	-
Defect energy spread [eV]	1	0.3	0.3	-	-

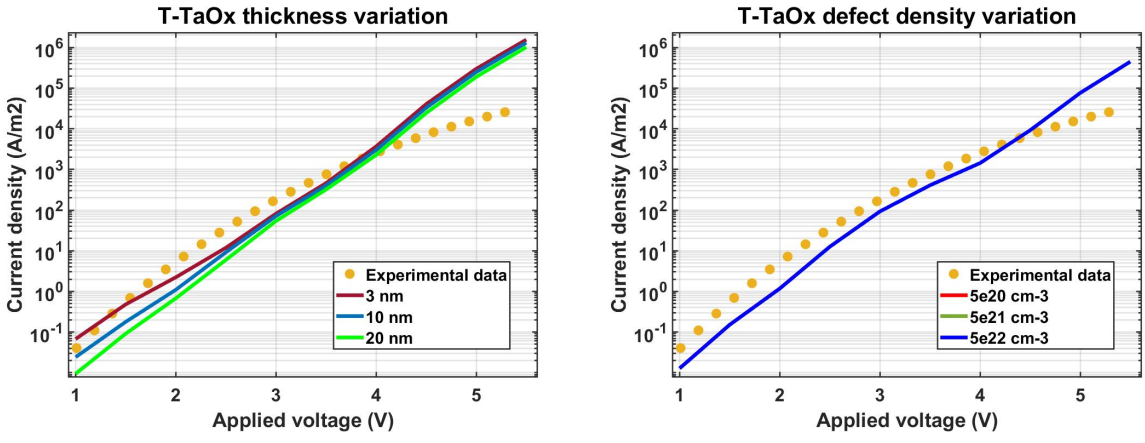
Table 4.1: Table of the tri-layer parameters.

4.2.1.2 T-TaO$_x$ parameter analysis

To check the consistency of the proposed model, two designs of experiment have been performed to study how the parameters of the thick TaO$_x$ layer affect the current density of the simulated cell. For what concerns the physical options of the simulation (see section 2.3.2), a DC sweep from 1 to 4 V is applied at the top electrode while the bottom electrode voltage is kept constant at 0 V. This is because the measurements below 1 V are inaccurate, as the pristine current is shadowed by the noise of the measurement. The temperature is set at 300 K, all the conduction mechanisms are activated and the simulation is performed with the defect density approach.

Fig. 4.4 (a) shows the output of the simulation for a thickness of the top TaO$_x$ layer of 3, 10 and 20 nm. The major difference is at low voltages, where direct tunneling (see section 2.2) plays a bigger role, hence the larger current for the thinner layer. At higher voltages the thickness has a negligible impact because of the high dielectric constant of the TaO$_x$ layer, since the voltage drop on the layer is close to null, as shown in fig. 4.5.

Fig. 4.4 (b) shows the output of the simulation for a defect density of the top TaO$_x$ layer of 5*10^{20}, 5*10^{21} and 5*10^{22} cm^{-3}. The simulation curves overlap as the defect density has no impact on the conduction of the TaO$_x$ layer. The band flattening effect of the high dielectric permittivity is dominating over the trap assisted tunneling in T-TaO$_x$.



(a) (b)

Figure 4.4: Simulation of the effect of the variation of T-TaO$_x$ thickness (a) and T-TaO$_x$ defect density (b) (simulation curves overlap) on the pristine state current density compared to experimental data.

From the studies on thickness variation and defect density of the TaO_x layer, it is clear that the resistance of this layer in this model is negligible compared to the ones of the interface layer and the stoichiometric HfO_2. The idea is therefore to substitute this layer with an equivalent electrode and reducing the simulated device to a simplified bi-layer composed of only hafnium oxide and interface TaO_x.

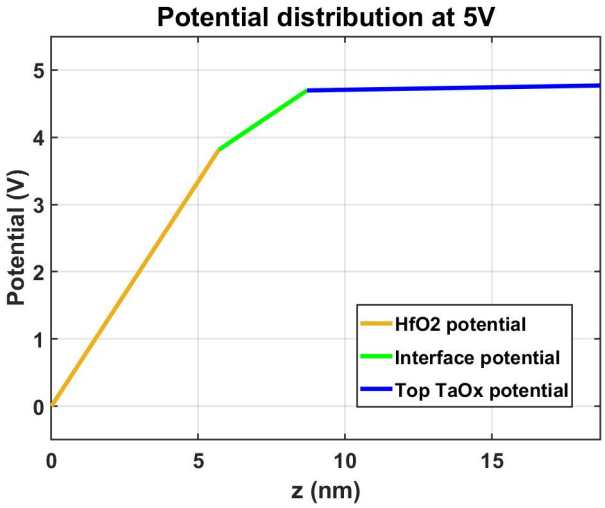


Figure 4.5: Potential distribution in the tri-layer for an applied voltage of 5 V.

4.2.2 Simple bi-layer simulations

Following the results of section 4.2.1, a new modeling approach of a bi-layer with an effective interfacial layer is proposed. The top TaO_x layer is now considered as an electrode with tweaked parameters, while the parameters of the interfacial layer have also been adjusted to match experimental data.

4.2.2.1 Device representation and parameters

A representation of the simulated device is shown in fig. 4.6, and a summary of the parameters is given in table 4.2. The lateral dimensions of the cell and the thicknesses of the HfO_2 layer and of the effective interface are kept as in the three-layer device. The parameters of the top electrode have been adjusted to fit the experimental data, although they differ from the standard values of both Ta_2O_5 and metallic tantalum. The bottom electrode is the default titanium nitride of the Ginestra® material library.

For HfO_2 the defect density is increased from $3*10^{19}$ to 10^{20} cm^{-3} for z between 4.5 and 5.7

nm to represent the absorption of oxygen species of the substoichiometric interface from the hafnium oxide, a modeling approach that is typical of interfaces between a stoichiometric oxide and a metallic material [26]. The same approach is applied to the effective interface, where the defect density goes from 10^{20} to $2*10^{20}$ for z going from 6.7 to 8.7 nm, as the interface should tend to a higher defect density the closer it is to the top electrode (the substoichiometric T-TaO_x).

	HfO_2	Interface	TE	BE
Work Function [eV]	-	-	5.22	4.57
Density of states [m^{-3}/J]	-	-	$9*10^{40}$	$9.037*10^{45}$
Band Gap [eV]	5.8	4.05	-	-
Electron affinity [eV]	2.4	3.45	-	-
Dielectric permittivity	26	28	-	-
Thermal conductivity [$Wcm^{-1}K^{-1}$]	0.0035	0.01	-	-
Tunneling mass [m_0]	0.21	0.4	-	-
Defect density [cm^{-3}]	$3*10^{19}$, 10^{20}	10^{20}, $2*10^{20}$	-	-
Defect energy mean	2.1	1	-	-
Defect energy spread	1	0.3	-	-
E_{AG} [eV]	3	-	-	-
p_0 [eÅ]	4.2	-	-	-
E_{AR} [eV]	-	0.2	-	-
Oxygen ion E_{AD} along z [eV]	1.2	1	-	-
Oxygen ion E_{AD} along x,y [eV]	1.2	1.2	-	-

Table 4.2: Table of the simple bi-layer parameters.

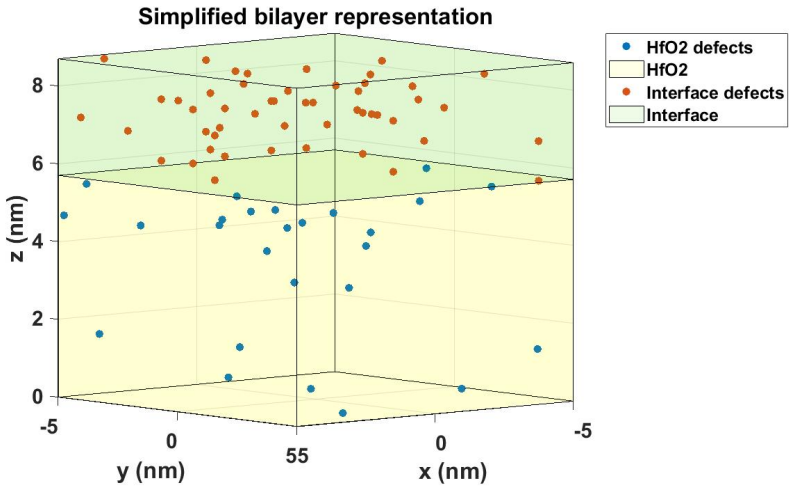


Figure 4.6: Simplified bi-layer cell 3D representation.

4.2.2.2 Pristine state study

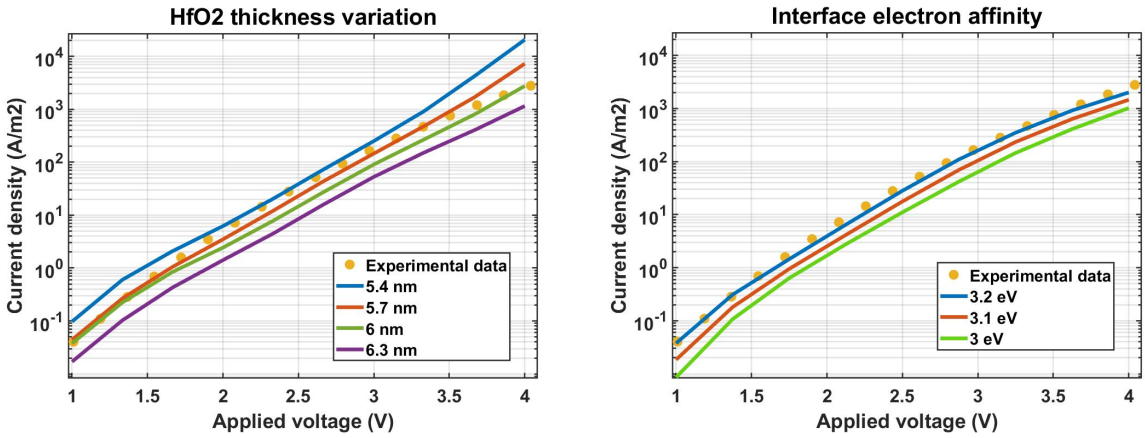
To check the consistency of the model, an analysis to verify the matching of process variations has to be performed. The model should be capable of capturing both the device to device variability and the change of processing steps, like the change of TaO_x nominal resistivity and HfO_2 thickness.

To do so, two designs of experiments have been performed by varying the simulated HfO_2 thickness and the interfacial TaO_x electron affinity. The output of the simulation is shown in fig. 4.7. The simulation conditions are the same as the ones used in section 4.2.1.2.

Fig. 4.7 (a) shows the simulated current density of a pristine simplified bi-layer for HfO_2 thickness of 5.4, 5.7, 6 and 6.3 nm. The result is a current increase for a thinner oxide layer, as expected since the tunneling process, both direct and trap-assisted, becomes easier for a thinner potential barrier. The curves cover the whole spectrum of the measured devices fabricated on the same chip, thus showing that this pristine model is capable of emulating the device to device variability intrinsic to the fabrication process.

Fig. 4.7 (b) shows the simulated current density of a pristine simplified bi-layer for interface electron affinity of 3, 3.1 and 3.2 eV. An increase of electron affinity is translated in a more conductive layer as the potential barrier of the conduction band is lowered. This simulation captures how a different conductivity of the interface can influence the current density of

the devices, and matches the experimental data gathered by measuring devices fabricated with different TaO_x resistivity values.



(a) (b)

Figure 4.7: Simulation of the effect of the variation of (a) HfO_2 thickness and (b) Interface electron affinity on the pristine state current density compared to experimental data.

4.2.2.3 Forming simulation

Once the pristine state model fits the experimental data, the next step in the modeling is the forming of the devices. Since forming is a time dependent process, the simulations are transients. As explained in section 2.1.2, the forming is modeled with a generation/recombination mechanism of oxygen ion/oxygen vacancy pairs. The beginning of the process is the electrostatic breakdown of the oxigen-metal molecules in HfO_2. Hafnium oxide is the most resistive material in the RRAM, so it is reasonable to assume that the majority of the voltage drops on this layer and the forming process begins in this region. The generation of oxygen vacancies causes a rapid increase of the current density in the region of their formation, increasing the temperature and subsequently the defect/ion generation rate. This positive feedback loop gives rise to very fast insulator-to-metal transitions during the forming of the devices.

Two parameters need to be tuned to match the experimental forming voltage: the generation activation energy E_{AG} and the polarizability p_0 (see section 2.1.2). Both these parameters have an exponential influence in the generation rate of defects in HfO_2. The energy reduction ΔE due to neighboring defects is set to 0. To match the experimental forming voltage of the devices, the polarizability is set to 4.2 eÅ and the activation energy

is set to 3 eV. These values are slightly different from literature, since usually E_{AG} is set to 2.7 eV and p_0 is 5.2 eÅ [26]. A higher activation energy and a lower polarizability both mean that the generation requires a higher electric field or a higher temperature to initiate the positive feedback of forming. There are two possible explanations for this parameter difference: first of all this is a model of a very complex system in which multiple simplifications have been introduced, first and foremost the reduction of the top TaO_x layer to an effective electrode. Another reason of this difference could be the fact that the values are extracted from devices in which the HfO_2 is sandwiched between two metals, while in this case one of the interfaces is with a substoichiometric oxide. Changing the nature of the interface could have an effect on the cristallinity of the HfO_2, thus altering the electric field required for the oxide breakdown.

For what concerns the simulation options of section 2.3, a 2.6 s voltage sweep is applied at the top electrode starting from 3 V and ending at 5.6 V, with a voltage ramp of 1 V/s. The initial temperature is set to 300 K and all the conduction mechanisms are activated. The charge of the generated oxygen ions is taken into account in the solution of the Poisson equation, and the Fourier equation solution is activated to capture the positive feedback effect. Also the material degradation is activated to capture the change in the potential distribution as the metallic filament is formed.

The usual filament size is estimated by considering a metallic hafnium filament between two metallic electrodes and is usually between 10x10 and 14x14 nm^2 [27], but the estimation in this case cannot be performed as the two layers of TaO_x play an important role in the post-forming resistance. The filament dimension in the devices is therefore approximated to 12x12 nm^2. This filament size would require a very large simulated cell and extremely long simulation times, therefore to limit these problems the generation is confined in a 2x2 nm^2 area at the center of the cell in the HfO_2 layer. The pre-forming data is scaled down to an area of 10x10 nm^2 from 30x30 μm^2, while the post-forming experimental current is scaled down by a factor of (2x2)/(12x12)=1/36. This scaling approach can be seen as simulating one fraction of a large structure in which multiple 2x2 nm^2 filaments would be put together in parallel. An example of a 4x downscaling is shown in fig. 4.8. A series resistance of 400 kΩ is added in the simulation. In the experimental forming, the series resistance used is 10 kΩ but since the filament was scaled to have a resistance 36 times greater, also the series resistance has to match the same scaling in the simulation.

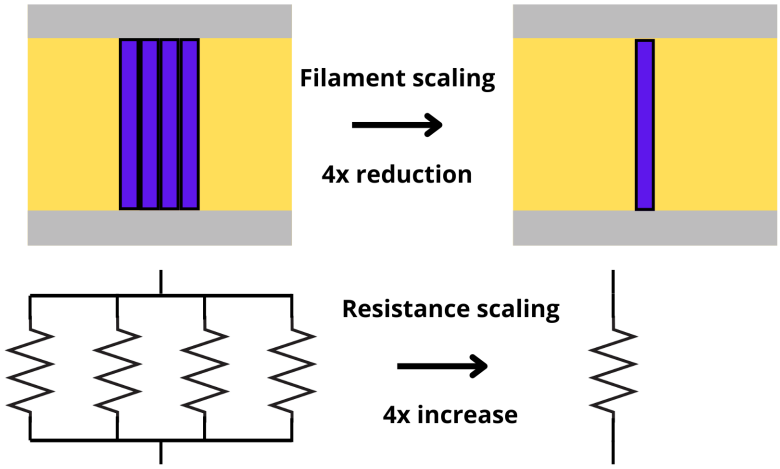
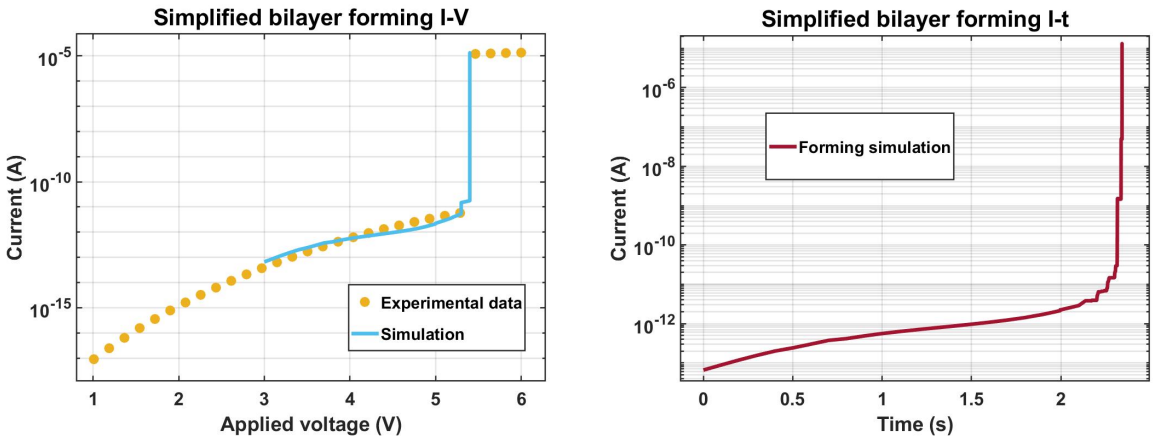


Figure 4.8: Example of a 4x filament downscaling.

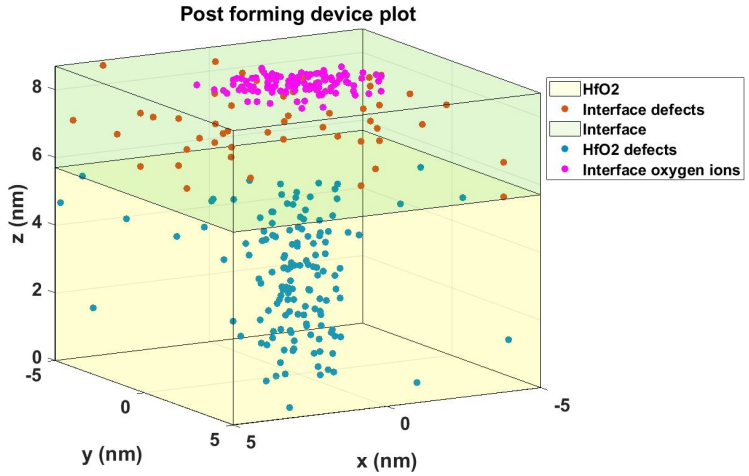
The current output of the forming simulation is shown in fig. 4.9. Fig. 4.9 (a) shows the I-V curve of the simulated device, demonstrating that the experimental forming voltage is captured by the model. The simulation was stopped when the post forming current had been reached. Fig. 4.9 (b) shows the I-t curve of the simulation, showing how the forming is a very sharp current spike in the device as the positive feedback effect of temperature increase and current increase makes the phenomenon very quick [28].



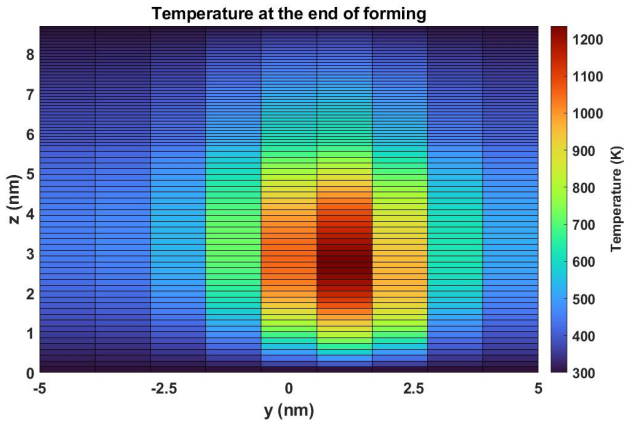
(a) (b)

Figure 4.9: Simulated forming curves of the simplified bi-layer as a function of applied voltage (a) and time (b).

Other relevant simulation outputs are shown in fig. 4.10. Fig. 4.10 (a) shows the simulated device at the end of the forming simulation. A filament of oxygen vacancies was formed at the center of the device, while the oxygen ions have drifted towards the top electrode due to the applied voltage. Fig. 4.10 (b) shows the temperature profile of a cut section of the cell through the axis of the filament. The maximum temperature reached is above 1220 K, much higher than literature values [29], and also if the simulation was not stopped once the post forming experimental current was reached the temperature was expected to increase even further since there are no self-mitigating effects to the positive feedback loop except for the series resistance.



(a)



(b)

Figure 4.10: Device plot at the end of the forming (c). Temperature profile of a cut section along the filament at the end of the forming simulation (d).

The forming study allowed to identify the flaw of this modeling approach, as it is too simple to capture the forming of the devices. The absence of an effective interruption of the positive current-temperature feedback of the forming leads to extremely high simulation temperatures, which is not physically plausible as the materials of the stack would melt and cause the failure of the devices. The solution to this problem is the introduction of a model that is halfway between the simplified bi-layer and the tri-layer of section 4.2.1.

4.2.3 Substoichiometric interface bi-layer simulations

The third and most successful model consists of a bi-layer in which the effective interface is rich of defects and the top electrode is considered as metallic tantalum.

4.2.3.1 Device representation and parameters

A representation of the simulated device is shown in fig. 4.11 and a summary of the relevant parameters is given in table 4.3.

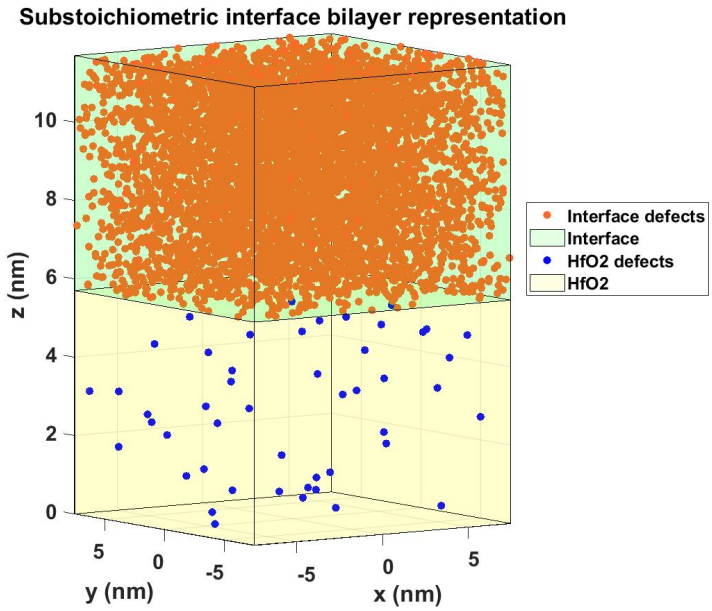


Figure 4.11: Substoichiometric interface bi-layer cell 3D representation.

The lateral dimensions of the cell are increased to 7.5x7.5 nm^2 ensure a proper heat calculation during the forming process. The thickness of the HfO_2 layer is kept to 5.7 nm, while the thickness of the effective interface is increased to 6 nm. To balance the thickness

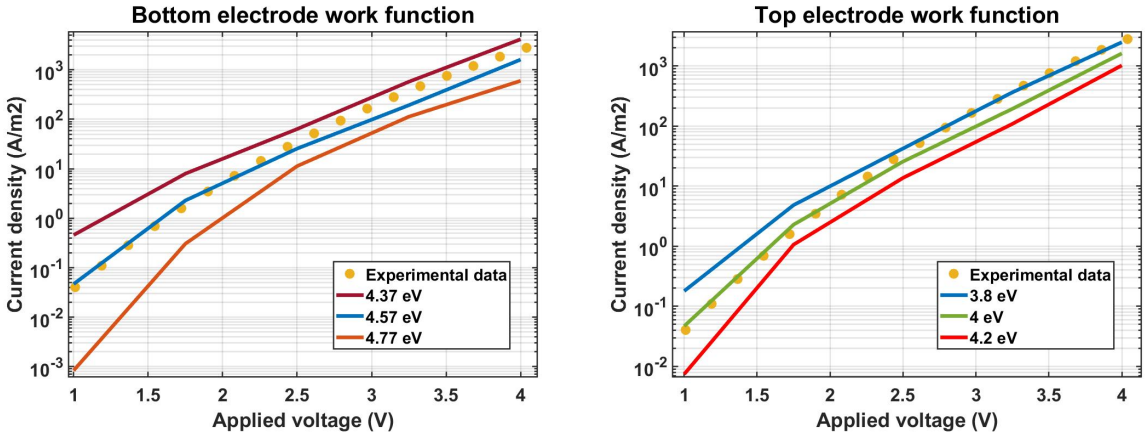
increase, the defect density has been increased aswell to a linear distribution from 5.7 to 11.7 nm along z going from 3*10^{21} to 6*10^{21} cm^{-3}, to maintain roughly the same overall conductivity of the layer. The thermal conductivity of the interface is set to 0.01 W cm^{-1} K^{-1} [30], and the thermal conductivity of HfO_2 is set to 0.0035 W cm^{-1} K^{-1} from previous models. The HfO_2 defect density is also a linear distribution from 0 to 5.7 nm going from 3*10^{19} to 5*10^{19}. This distribution is a better match to literature values of stoichiometric samples [24]. The top electrode work function now matches literature values of tantalum, but the density of states has been lowered from 9*10^{45} to 9*10^{40} m^{-3}/J to model the fact that the layer is substoichiometric TaO_x. The bottom electrode is the default titanium nitride of the Ginestra® material library.

	HfO_2	Interface	TE	BE
Work Function [eV]	-	-	4	4.57
Density of states [m^{-3}/J]	-	-	9*10^{40}	9.037*10^{45}
Band Gap [eV]	5.8	4.05	-	-
Electron affinity [eV]	2.2	3.2	-	-
Dielectric permittivity	22	25	-	-
Thermal conductivity [$Wcm^{-1}K^{-1}$]	0.0035	0.01	-	-
Tunneling mass [m_0]	0.3	0.3	-	-
Defect density [cm^{-3}]	From 3*10^{19} to 5*10^{19}	From 3*10^{21} to 6*10^{21}	-	-
Defect energy mean	1.65	1	-	-
Defect energy spread	1.1	0.4	-	-
E_{AG} [eV]	3.2	-	-	-
p_0 [eÅ]	4.2	-	-	-
E_{AR} [eV]	-	0.2	-	-
Oxygen ion E_{AD} along z [eV]	1.2	0.9	-	-
Oxygen ion E_{AD} along x,y [eV]	1.2	1.2	-	-

Table 4.3: Table of the substoichiometric interface bi-layer parameters.

4.2.3.2 Pristine state study

An extensive analysis of the main parameters affecting the conduction of the pristine cell has been performed in order to identify possible process improvements to obtain devices with better performance. In all tests, a 3 s transient sweep is performed with a voltage ramp of 1 V/s from 1 to 4 V. The temperature is fixed at 300 K and all conduction mechanisms are activated. The charge of occupied traps is neglected in the potential calculation, as well as the heat equation. Since the defect density of the interface is close to the critical value of 10^{22} cm^{-3}, degradation is activated to capture a local change of the dielectric permittivity. The first analysis is the effect of the variation of the electrode work functions, and the result of the two designs of experiment is shown in fig. 4.12.



(a) (b)

Figure 4.12: Simulation of the effect of the variation of (a) bottom electrode work function and (b) top electrode work function on the pristine state current density compared to experimental data.

Fig. 4.12 (a) shows the pristine current density for a bottom electrode work function of 4.37, 4.57 and 4.77 eV. The current increases as the work function decreases over the whole sweep. This can be explained by the fact that a lower work function means a higher Fermi level of the metal, so there are more electrons available to tunnel through HfO_2 and participate to the conduction, hence increasing the current.

Fig. 4.12 (b) shows the pristine current density for a top electrode work function of 3.8, 4 and 4.2 eV. The current increases for a decrease of work function at all voltages. This can be explained by the alignment of interface conduction band and defect level with the Fermi

level of the metal, as a better alignment favors a higher conduction.

These results show that the tuning of the processing of either the bottom electrode or the TaO_x can achieve a change of current of the pristine devices.

The second analysis performed is on the relevant parameters of the hafnium oxide layer, and the output of the simulations is shown in fig. 4.13.

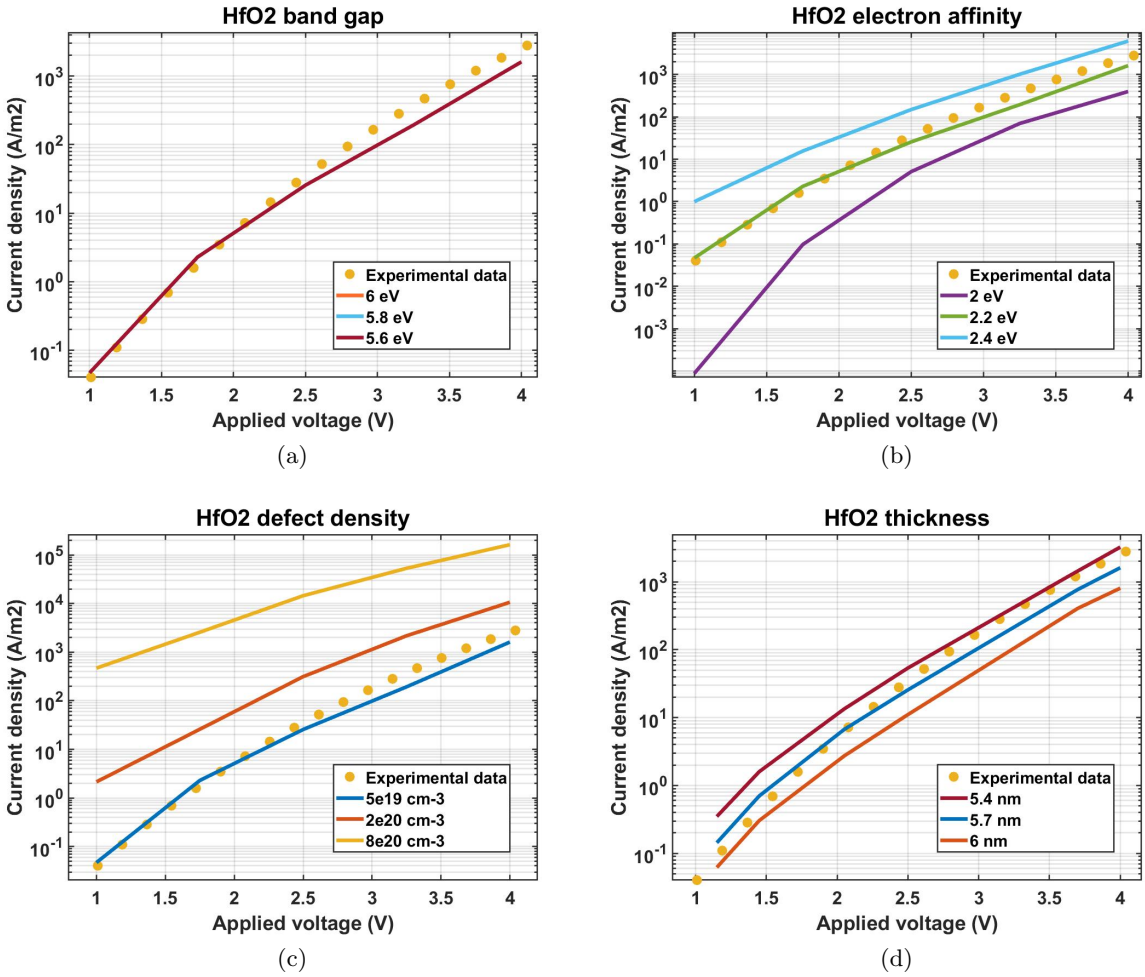


Figure 4.13: Simulation of the effect of the variation of hafnium oxide (a) band gap (simulation curves overlap), (b) electron affinity, (c) defect density and (d) thickness on the pristine state current density compared to experimental data.

Fig. 4.13 (a) shows the pristine current density of the devices for band gaps of 5.6, 5.8 and 6 eV. Since the band gap is very large, its variation has no impact on the overall conduction, because the position of the conduction band is determined by the electron affinity and the

variation of bad gap is too small to allow for efficient drift-diffusion conduction mechanisms typical of semiconductors with smaller band gaps.

Fig. 4.13 (b) shows the pristine current density of the devices for electron affinities of 2, 2.2 and 2.4 eV. An increase of electron affinity causes a large increase of current density. This is because the electron affinity has two major effects: improving the alignment of the trap energy level of hafnium oxide and the bottom electrode and also reducing the potential barrier represented by the conduction band of hafnium oxide.

Fig. 4.13 (c) shows the pristine current density of the devices for peak defect densities of $6*10^{19}$, $2*10^{20}$ and $8*10^{20}$ cm^{-3}. It is clear that a defect density increase causes a current increase. The defect density has a direct correlation with the conductivity of a layer, therefore a more defective hafnium oxide favors larger currents. This means that a process improvement to get more conductive devices could be the deposition of hafnium oxide in an environment with less oxygen.

Fig. 4.13 (d) shows the pristine current density of the devices for thicknesses of 5.4, 5.7 and 6 nm. An increase of the thickness causes a reduction of the current, as the addition of more material in the stack could be seen as adding a series resistance. This graph is particularly important since these 3 simulation curves cover the whole spectrum of the pristine current densities of the measured devices, which means that a device to device variation is captured by varying the hafnium oxide thickness of less than 0.6 nm. This result is promising for the correctness of the model, as what dictates the variation between each device is the roughness of the interfaces between each layer, which can be correlated to the effective thickness of the layers in the devices.

The last analysis performed on the pristine device is on the parameters of the effective interface to check the consistency of the model, and the output of the simulations are shown in fig. 4.14.

Fig. 4.14 (a) and (b) show the pristine current density of the devices for band gaps of 3.9, 4.05 and 4.2 eV and for electron affinities of 3, 3.2 and 3.4 eV respectively. The influence of these parameters has no impact on the conductivity of the layer. This is because in the Ginestra® simulation framework once an electron reaches the conduction band of an oxide it is assumed that it will reach the electrodes. Therefore once an electron manages to tunnel through the hafnium oxide into the conduction band of the interface it will automatically reach the electrode, neglecting the rest of the energy landscape. This means that the position of the bottom of the conduction band or the position of the top of the valence

band of the effective interface has no impact on the overall conduction as long as electrons that travel through hafnium oxide end up in the conduction band.

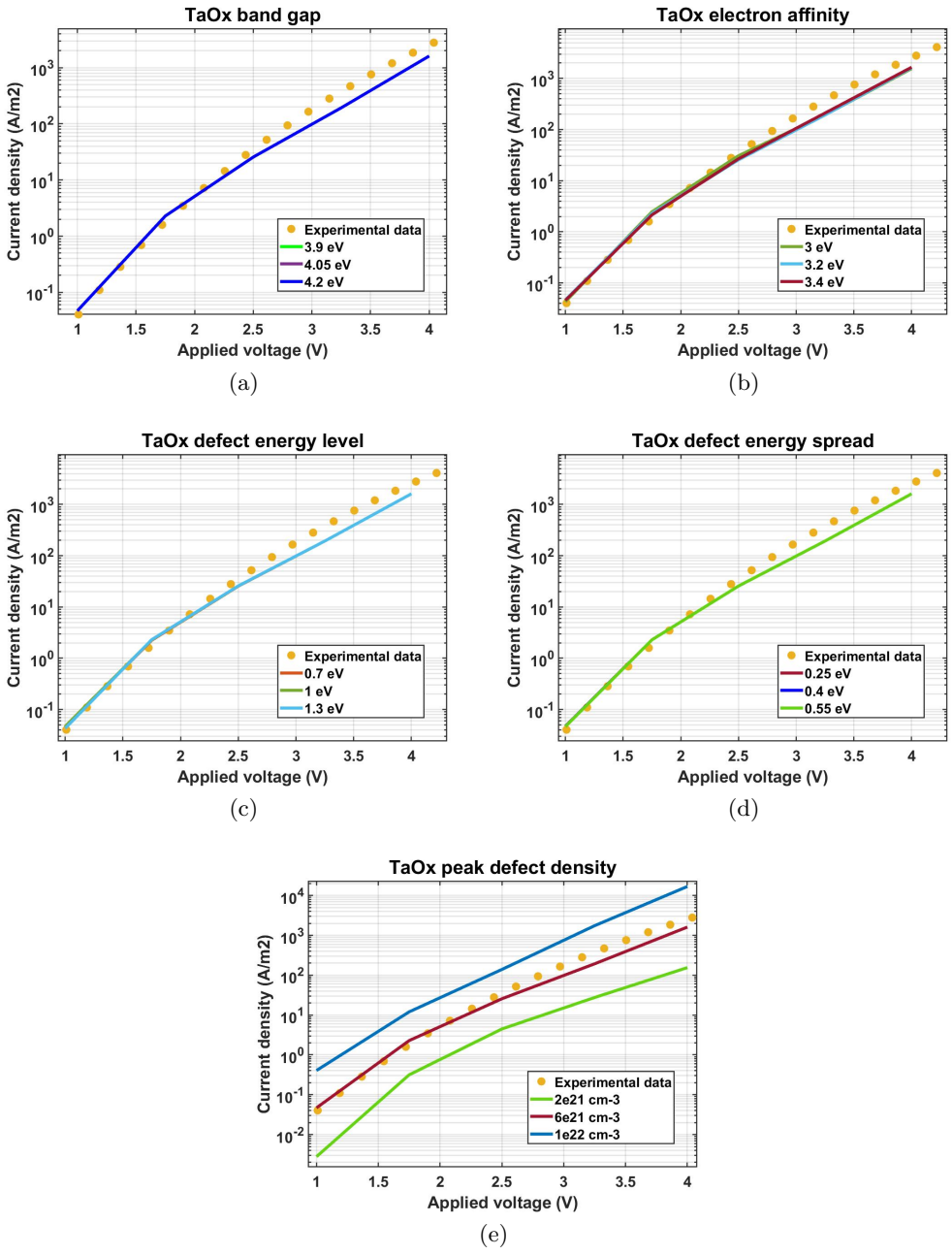


Figure 4.14: Simulation of the effect of the variation of substoichiometric tantalum oxide (a) band gap (simulation curves overlap), (b) electron affinity, (c) peak defect density, (d) defect energy level (simulation curves overlap) and (e) defect energy spread (simulation curves overlap) on the pristine state current density compared to experimental data.

Fig. 4.14 (c) and (d) show the pristine current density of the devices for the variation of the energy distribution of the defects of the effective interface. Neither the defect energy level mean (c) nor the defect energy spread (d) have an impact on the overall conduction for the same reason as to why band gap and electron affinity have no impact. The energy landscape is irrelevant for the simulation tool once an electron manages to reach the conduction band. Fig. 4.14 (e) shows the pristine current density of the devices for peak defect densities of $2*10^{21}$, $6*10^{21}$ and 10^{22} cm^{-3}. An increase of defect density in the effective interface causes an increase of potential drop on the hafnium oxide layer, as the dielectric constant of the interface tends to increase. A schematic of this effect is shown in fig. 4.15. Since the HfO_2 layer is the current bottleneck, a larger voltage drop is translated in an increase of the current.

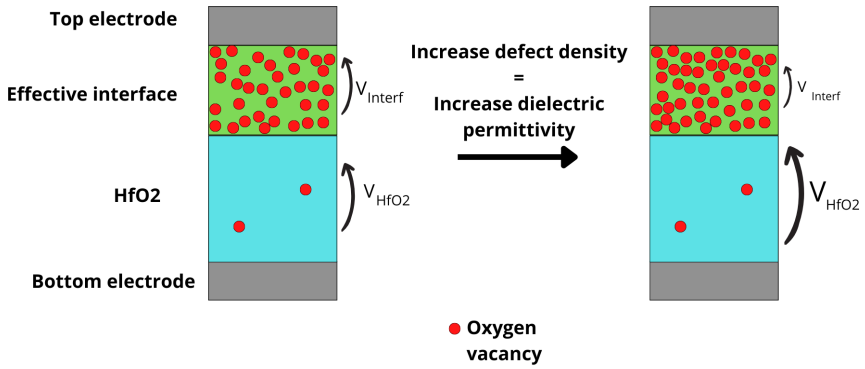


Figure 4.15: Effect of the defect density increase on the voltage partition in the bi-layer.

4.2.3.3 Forming simulation

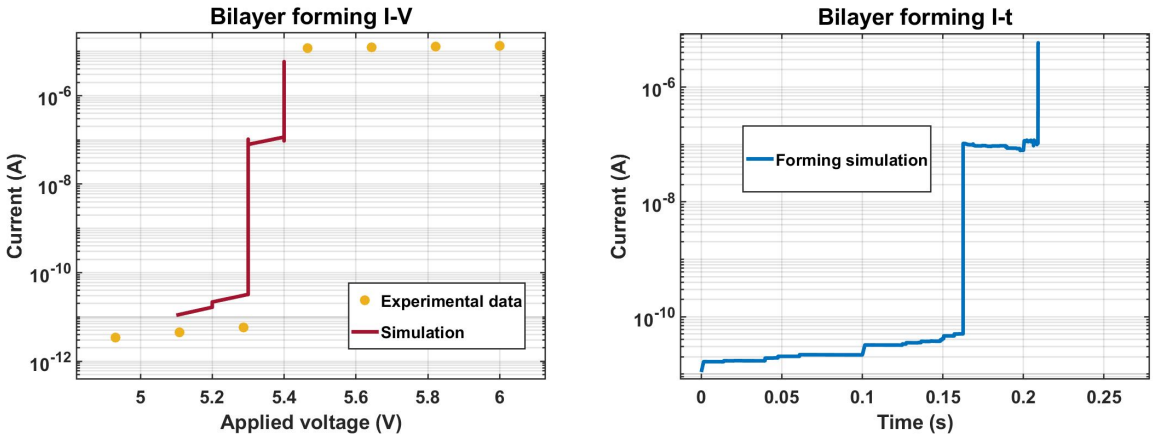
After the analysis of the pristine state parameters, the next step is to model the forming of the device. The energy reduction ΔE due to neighboring defects is set to 0, and to match the experimental forming voltage of the devices the polarizability is set to 4.2 eÅ and the activation energy is set to 3.2 eV. The values are once again different from literature values but the explanation given in section 4.2.2.3 is still valid. In addition to generation, the recombination of oxygen ions and oxygen vacancies is activated in the effective interface to capture the oxygen reservoir effect of the tantalum oxide, as it should absorb the oxygen released from the hafnium oxide..

For what concerns the simulation options of section 2.3, a 0.5 s voltage sweep is applied at

the top electrode starting from 5.1 V and ending at 5.6 V, with a voltage ramp of 1 V/s. This short sweep was used to reduce simulation times, as the large number of defects in the effective interface causes a huge computational load and the generation/recombination below 5.1 V is negligible compared to the one during the forming process. A series resistance of 400 kΩ is added like in the case of the simplified bi-layer. The initial temperature is set to 300 K and all the conduction mechanisms are activated. The charge of the generated oxygen ions is taken into account in the solution of the Poisson equation, and the Fourier equation solution is activated to capture the positive feedback effect. The material degradation is activated to capture the change in the potential distribution as the metallic filament is formed.

Once again the generation is limited in a 2x2 nm^2 region at the center of the cell, to limit simulation times and to ensure a proper heat distribution during forming. The same explanation of section 4.2.2.3 applies for this simulation.

The current output of the forming simulation is summarized in fig. 4.16.

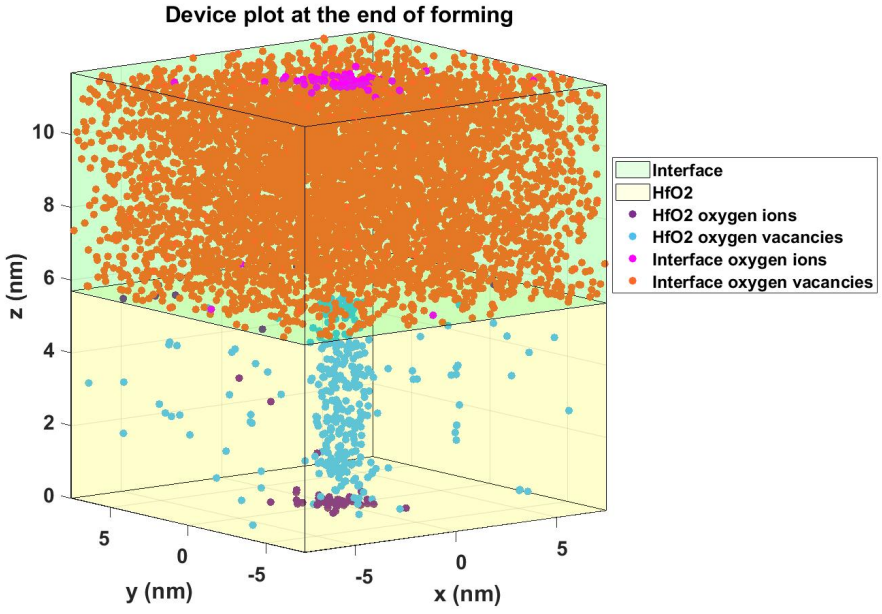


(a) (b)

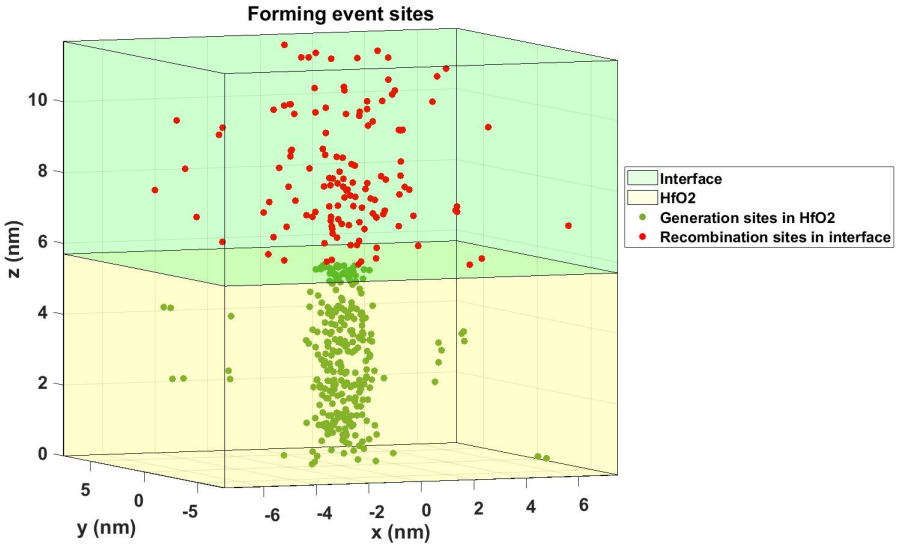
Figure 4.16: Simulated forming curves of the bi-layer as a function of applied voltage (a) and time (b).

Fig. 4.16 (a) shows the I-V curve of the simulated device, showing that the forming voltage is matched and the post forming current levels can be reached. 4.16 (b) shows the time evolution of the simulated current. The first current spike is caused by the generation of a single defect in a peculiar position both in the 3D space and in the energy space, generating a very efficient conduction path. The second spike is caused by the positive feedback loop

of joule heating and defect generation. The simulation was interrupted before finishing the sweep due to an incredibly heavy computational and memory load.
The 3D device at the end of the forming simulation is shown in 4.17 (a).



(a)



(b)

Figure 4.17: Device plot of the device at the end of the forming (a). Generation and recombination sites at the end of forming (b).

It is clear that a filament was generated in the center of the cell and a large number of oxygen ions has migrated to the top electrode due to the applied voltage.
To better represent the interplay between generation of defects in hafnium oxide and recombination in the effective interface, fig. 4.17 (b) shows a 3D representation of the generation and recombination sites of the two layers. The oxygen ions preferably move along the z axis, but they spread due to Coulomb repulsion between each other and local variations of the electric field. This spread gives rise to the mushroom-like shape of the structure, as the oxidized area can be seen as a hemisphere with the centre corresponding to the tip of the filament. The oxidation decreases for an increasing distance from the tip of the filament.
Again from 4.17 (a) an accumulation of oxygen ions can be seen at the base of the filament in proximity of the bottom electrode. This is due to the fact that the bottom electrode temperature is fixed at 300 K as shown in the temperature profile along the filament axis at the end of the simulation in fig. 4.18, therefore the movement of the species in that region is much slower compared to the ions generated in the region at high temperature (see section 2.1.2, eq. 2.1).

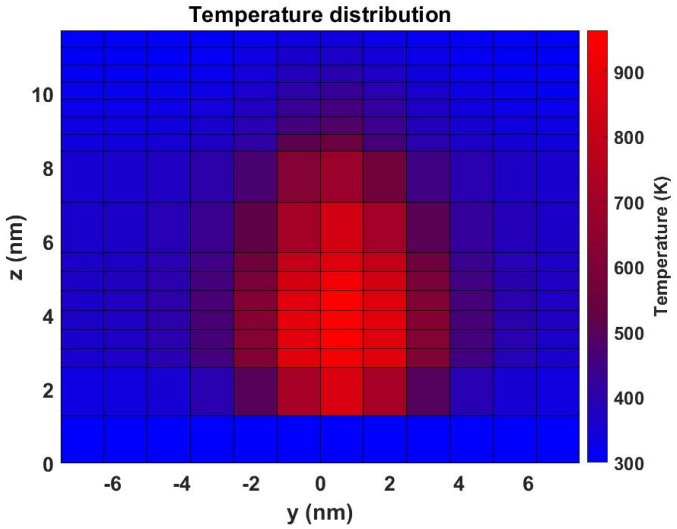


Figure 4.18: Temperature at the end of the forming in a cut section along the filament.

Always from 4.18 the thermal effect of the effective interface can be seen. The low thermal conductivity of the layer causes heat confinement in the region where the filament is forming, thus enhancing the forming process. The maximum temperature is in line with literature values [29].

Fig. 4.19 shows how the potential changes from the beginning to the end of the simulation along the filament axis. It is clear that the voltage drop on the filament is greatly reduced due to the high defect density reached of roughly $8*10^{21}$ cm^{-3}, so it is more distributed between filament and effective interface at the end of the process. The overall voltage drop on the device is also greatly reduced due to the presence of the series resistance, which together with the oxidation of the interface acts as an interruption to the positive feedback of the forming.

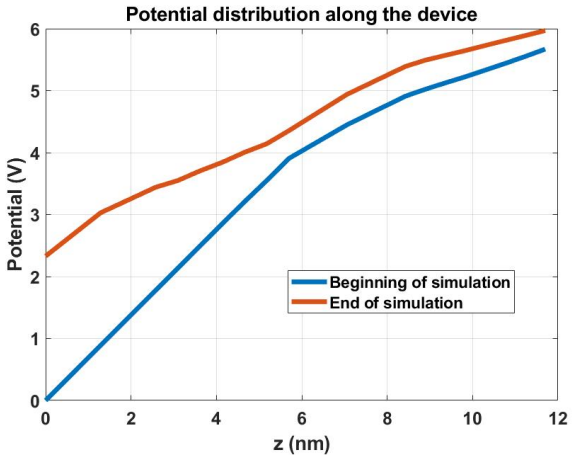


Figure 4.19: Comparison of the potential distribution along the filament axis at the beginning and at the end of the simulation.

4.2.4 Post-forming fit

Since the forming simulation was interrupted before reaching the end of the sweep, it was not possible to use the output as a post-forming model and a new one had to be built following the results of the simulation. The structure of the post forming model is shown in fig. 4.20, with a list of the parameters in table 4.4. The proposed model is a two-layer device where the HfO_2 layer thickness is 5.7 nm and it contains the filament, modeled as a 2x2 nm^2 region with a defect density of $8*10^{22}$ cm^{-3}, and an effective interface of 3 nm thickness in which the defect density has a normal distribution with peak density at the four top corners of the device of $10*22$ cm^{-3} and lowest value in proximity of the filament tip. The result is a clear region where the defect density is much lower, as can be seen in the device plot of fig. 4.20, and this represents the oxidized area obtained in section 4.2.3.3. The cell size has been changed to 6x6 nm^2 to limit the analysis to the filament

surroundings and the dielectric constants of the layers have been adjusted to match the experimental data.

	HfO_2	Interface	TE	BE
Work Function [eV]	-	-	4.1	4.57
Density of states [m^{-3}/J]	-	-	$9*10^{40}$	$9.037*10^{45}$
Band Gap [eV]	5.8	3.6	-	-
Electron affinity [eV]	2.4	3.6	-	-
Dielectric permittivity	300	50	-	-
Thermal conductivity [W cm^{-1} K^{-1}]	0.005	0.01	-	-
Tunneling mass [m_0]	0.25	0.3	-	-
Defect density [cm^{-3}]	$8*10^{22}$	From 10^{20} to 10^{22}	-	-
Defect energy mean	1.55	0.6	-	-
Defect energy spread	1.1	0.4	-	-
E_{AG} [eV]	-	From 1.1 to 2.05	-	-
p_0 [eÅ]	-	10	-	-
E_{AR} [eV]	0.2	-	-	-
Oxygen ion E_{AD} along z [eV]	0.8	0.9	-	-
Oxygen ion E_{AD} along x,y [eV]	0.8	1.2	-	-

Table 4.4: Table of the post forming device parameters.

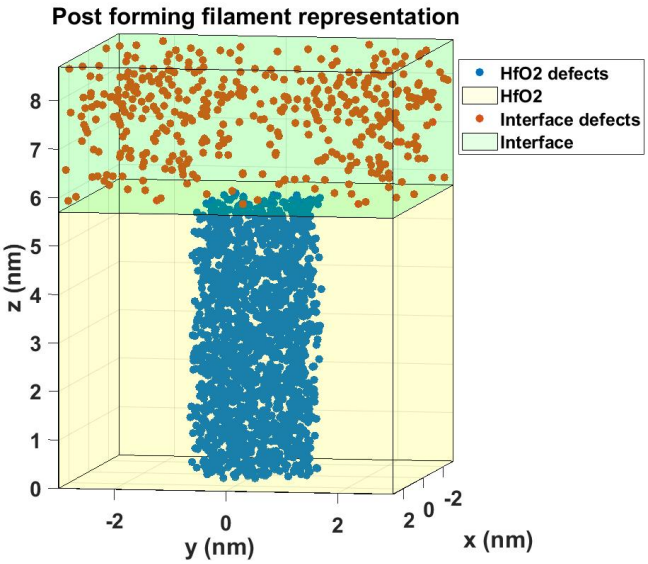


Figure 4.20: Post forming filament cell representation, showing the increasing defect density from filament tip to device edges.

A simulation of a positive and negative read operation performed on the cell is shown in fig. 4.21. The simulations consist of two DC sweeps from 0.01 to 0.25 and from -0.01 to -0.25 V applied at the top electrode with fixed temperature at 300 K and all conduction mechanisms enabled. The Fourier equation solution is neglected, as well as the effect of the charge of filled traps. The experimental data has been downscaled by a factor 36 for the same reason as explained in section 4.2.2.3.

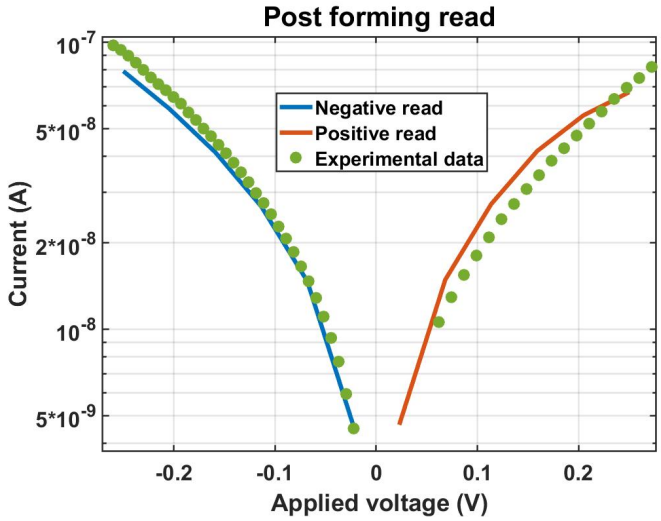


Figure 4.21: Simulated positive and negative read operation after filament formation.

Due to the extremely high defect density, the alignment of the sub-bands created by the defects is the key to the fit of the data. The energy levels of the traps of the two layers are aligned and there is a conduction sub-band from bottom electrode to effective top electrode.

4.2.5 Negative Set simulation

The next step is the modeling of the first resistance modulation of the cell, the negative set. The device is currently in a state of high resistance and the objective is to demonstrate that the mechanism for which the TiN/HfO_2/TaO_x/TiN stacks differ from standard RRAM devices is the interplay between HfO_2 and TaO_x in this model, which causes the set to be during a negative sweep instead of positive.

Four consecutive simulations are performed to capture the gradual nature of the process. The simulations are performed at a fixed temperature of 300 K with the voltage being applied at the top electrode with a ramp of 1 V/s. The charge of filled traps and of oxygen ions is neglected in the solution of the Poisson equation and the Fourier equation is also not enabled. Generation of oxygen ion/oxygen vacancy pairs is activated in the interface layer with activation energy E_{AG} gradually increasing from 1.1 to 2.05 eV and polarizability p_0 10 eÅ, while recombination is activated in hafnium oxide with activation energy E_{AR} 0.2 eV. The gradual increase of generation energy is introduced to match the experimental data, as the interface layer degrades at a slower rate at lower voltages until the current spikes due to a cascade of defects generation. The applied voltage sweeps are from -0.48 to -0.53 V, -0.54 to -0.6 V, -0.6 to -0.71 V and -0.71 to -0.75 V, although the final simulation was stopped at -0.72 V once the experimental peak was reached.

The I-V output of the simulation is shown in fig. 4.22, and the evolution of the device at the beginning and at the end of every step is shown in fig. 4.23.

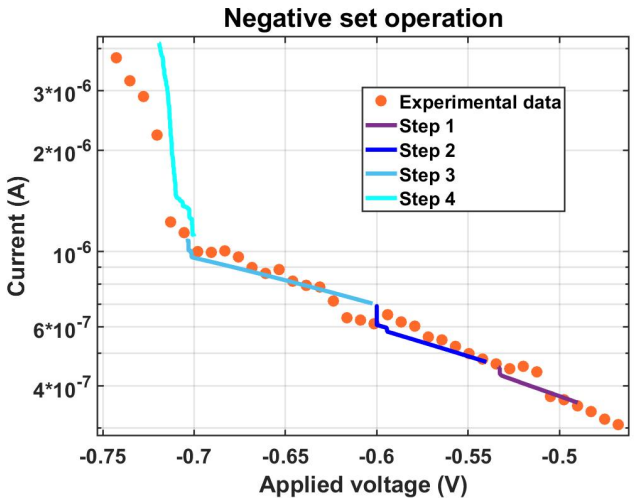


Figure 4.22: I-V output of the negative set simulation in 4 steps.

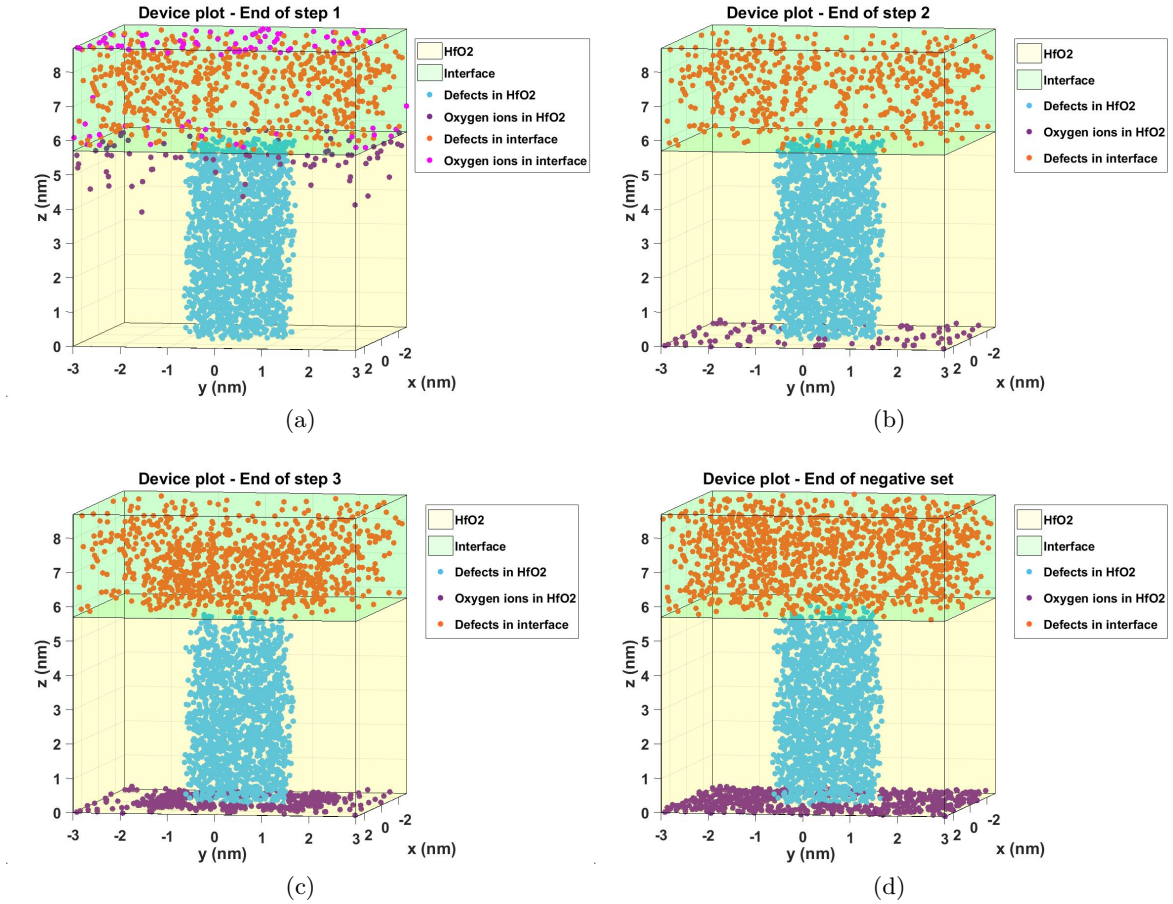


Figure 4.23: Device plot at the end of step 1 (a), step 2 (b), step 3 (c) and step 4 (d).

To highlight the interplay between the two layers, fig. 4.24 shows the recombination sites in hafnium oxide and the generation sites in the interface layer. It is clear that the high concentration of generated defects in the interface is matched by the erosion of the filament tip. This means that both the resistances of the two structures change during the process and that the decrease of resistance of the interface has a bigger impact on the current than the increase of resistance due to the filament being shortened of around 0.3 nm. This means that in the proposed model the oxidized region of tantalum oxide is the dominating element of the devices and its modulation is what dictates the state of the devices, which is a novelty compared to standard RRAM devices.

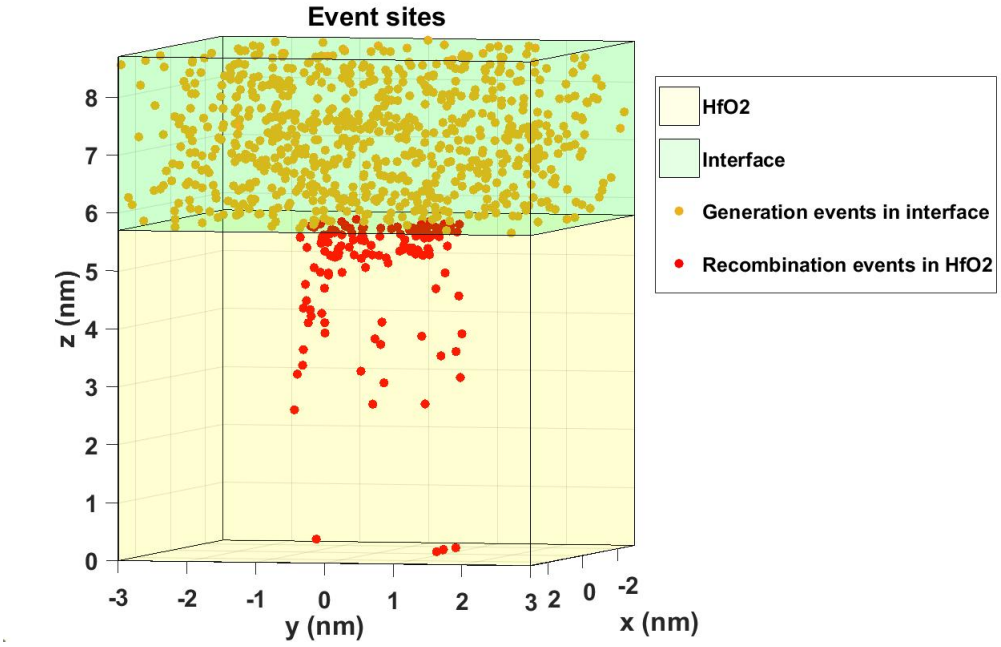


Figure 4.24: Plot of generation and recombination events during the SET operation.

Another important detail is the migration of oxygen ions towards the bottom electrode in this proposed model, which could be an explanation to the fact that after the devices in high resistance state are switched to LRS it is never possible to reach again the same high resistance state. The loss of species during switching could also explain the failure of the devices, as multiple consequent cycles would cause the active volume to spread and lead to a non-functioning device. The migration of oxygen ions could also mean that an oxidation process happens also at the bottom electrode, in which case it would require to adjust its parameters in the following simulations.

To conclude the analysis of the set simulation, a negative sweep from -0.75 to -0.1 V and positive sweep from 0.1 to 0.75 V is performed to check if the final device matches experimental data. The temperature is fixed at 300 K, all conduction mechanisms are activated and the same simplifications used for the negative set are applied here.
The output of the simulation is shown in fig. 4.25. It is clear that although the negative sweep fits the experimental data, the positive sweep only matches the experimental values at low voltages or at the end of the sweep. This could be explained by the fact that the current is fully controlled by trap assisted tunneling, and due to the initial difference in electrode work functions the presence of a built-in field makes the positive and negative sweeps asymmetric by facilitating the current for negative voltages, and reducing it for positive voltages.

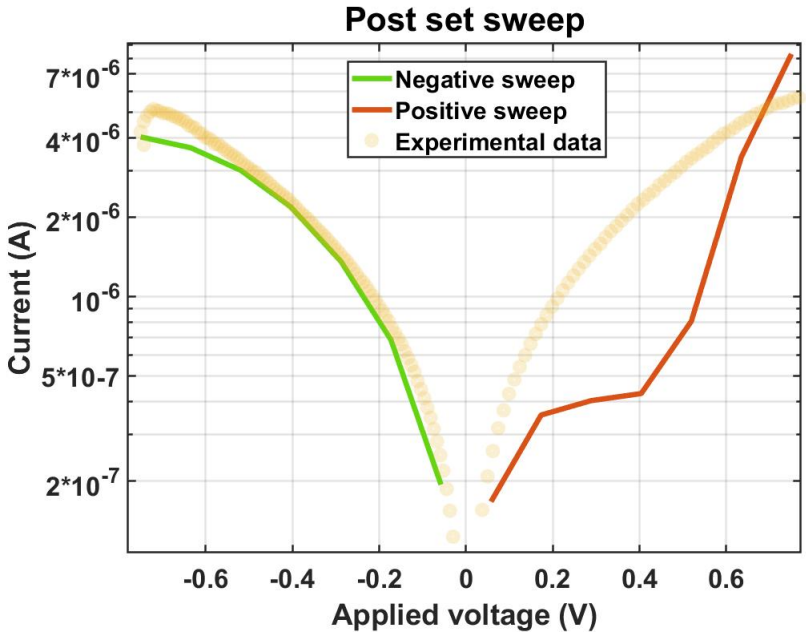


Figure 4.25: Simulated sweep after performing negative set operation.

Chapter 5

Conclusions and Outlook

5.1 Conclusions

In this proposed work a model describing RRAM devices based on HfO_2 and substoichiometric TaO_x has been developed, and a set of parameters has been found to match experimental data. By using the Ginestra® simulation tool, the proposed physical mechanisms dictating the behaviour of the devices are explained, showing how they can affect the forming and switching processes.

The modeling posed numerous challenges, from the complexity of the simulation tool to creating a model capable of capturing pristine state, forming and negative set, and the learnings and results are summarized as follows:

- From Capacitance-Voltage measurements it was possible to establish that the bulky TaO_x layer has all the characteristics of a metal and can be considered as one in the modeling, thus simplifying the representation of the devices in the simulation environment. To compensate for this structural change, the parameters of the interface TaO_x and of HfO_2 as well as of the top electrode have to be changed from standard values.
- The pristine state devices can be modeled as a bilayer composed of HfO_2 and an effective interface rich in oxygen vacancies. This model is capable of capturing the device to device variability by considering how the roughness affects the effective thickness of the fabricated devices, while also capturing the change of processing conditions that results in a different resistivity of the TaO_x and of the HfO_2 layers.

- The forming of the devices can be modeled with generation of oxygen vacancies in HfO_2, oxygen ion migration and recombination of oxygen ions and oxygen vacancies in the effective interface. The effect of Joule heating plays a critical role in the process, as it affects the mobility rate of oxygen ions and the generation rate of oxygen vacancies. The resulting structure is a filament with a hemispherical oxidation region in the effective interface.
- The negative set can be modeled with generation of oxygen vacancies in the effective interface, oxygen ion migration and recombination in the filament region. After forming, the device is in HRS and the most resistive part of the device is the oxidized region in the effective interface. While the filament resistance increases due to the erosion of the filament tip, the increased defect density of the effective interface causes this layer to become way more conductive, thus reducing the overall resistance of the device.
- The outcome of the modeled set does not match the post set experimental data, suggesting that the model lacks some details to capture the next switching operation to perform. This could be tied to the chosen parameters, to the device structure, to the simulation conditions or to other phenomena that has not been taken into consideration, like the oxidation of the bottom electrode.

5.2 Outlook

This master thesis work was focused on the modeling of $TiN/HfO_2/TaO_x/TiN$ devices on Ginestra®. The proposed modeling shows a new, exciting approach on how to emulate complex devices on a Kinetic Monte Carlo simulator. The introduction of an effective interface approach to substitute substoichiometric layers not only effectively captures the key mechanisms of the devices, but also helps in drastically improving the simulation times. The extension of this approach to other types of bilayers could prove very helpful in building a more consistent model capable of simulating any combination of oxide plus conductive metal oxide bilayer.

Before extending this approach to different technologies, however, a refinement of the simulation conditions and the device representation is required. A better forming and switching fit would require a more attentive study on the temperature and the multiple energy parameters of the devices in the various operations. The measurement of other parameters of

the devices like oxide stoichiometry, electrode work function or other relevant parameters could also be used to improve the modeling and obtain more accurate simulations.
The model could be used to improve the fabrication process and obtain devices with the required specifications. The reduction of the forming voltage for CMOS integration is one of the key aspects of the development of RRAM technologies. I-V symmetry, linearity, stochasticity and number of available resistance levels of the devices could be studied thanks to the model, as well as the retention capability of the devices.
Aspects like the effect of the environment temperature and the generation of free carriers by illumination could be explored in order to develop improved forming and switching procedures. The simulation tool allows to explore also the behaviour of the cell with a series transistor to emulate the 1T1R configuration used in crossbar arrays, thus allowing to check the performance of the devices in the crossbar array implementation.

Appendix A

Supplementary information on material characterization

A.1 Thermal conductivity measurement

This chapter is dedicated to the results of the thermal conductivity measurements following the method of section 3.2.

To extract the thermal conductivity, multiple devices with varying dimensions have been fabricated by deposition of the layers of the RRAM stack and subsequently processed by lithographic etching to obtain the desired shape for a 4 probe measurement. Since the most relevant parameter for forming simulations is the thermal conductivity of HfO_2, 3 chips were fabricated with HfO_2 thickness of 6, 12 and 18 nm, while all the other layer thicknesses are kept constant. The layout of the fabricated devices is shown in fig. A.1. The 4 large areas dedicated to each structure are for the 4 probes, and the electrical resistance of the thin lines is measured. The indicated dimensions are in μm and refer to the thin lines. Multiple structures with the same dimensions are available to perform statistical analysis. The idea is to exploit the insulating nature of the oxides of the stack and use the TiN top electrode as the metallic heater.

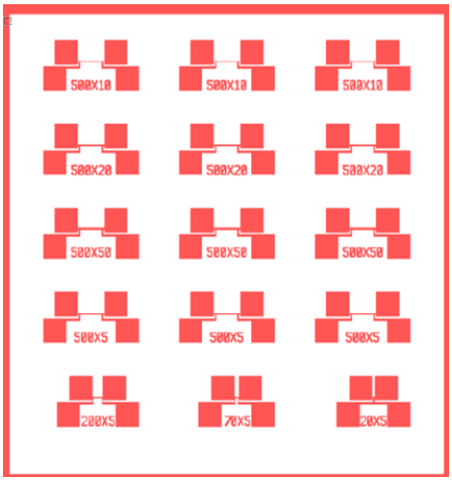
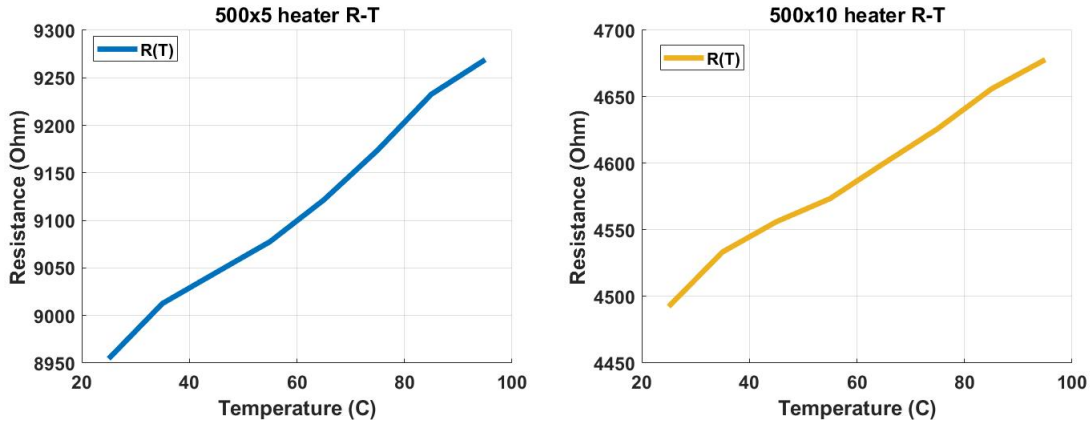


Figure A.1: Layout of the devices for the 3Ω measurement.

The relation between temperature and resistance of two thin metallic lines with different width is shown in fig. A.2. The resistance measurement is performed by applying a 0.2 V reading voltage and measuring the current, with device temperatures going from 298 K to 388 K with steps of 10 K. There is an expected linear increase of the TiN resistance with temperature, which is a good sign that indicates that the metal lines are functioning and can be used for the thermal measurements.



(a) (b)

Figure A.2: Resistance-Temperature measurements for heaters with dimensions 500x5 μm^2 (a) and 500x10 μm^2 (b).

The next step of the measurement is the heating of the structure by applying a DC voltage. The fabricated structures do not show any resistance change even at high voltages (e.g. 5

V). Increasing the voltage too much causes an irreversible failure of the structures, where the electrical resistance drops drastically and physical damage appears on the devices.

Bibliography

[1] K. Y. S. E. K. A.-S. S. F. Kavehei O., Iqbal A. and A. D., "The fourth element: characteristics, modelling and electromagnetic theory of the memristor," *Proc. R. Soc.*, vol. 466, pp. 2175–2202, 2010.

[2] C. de S. Dias and P. F. Butzen, "Memristors: A journey from material engineering to beyond von-neumann computing," *Journal of Integrated Circuits and Systems*, 2021.

[3] M. A. Zidan, J. P. Strachan, and W. D. Lu, "The future of electronics based on memristive systems," *Nature Electronics*, vol. 1, pp. 22–29, 2018.

[4] Y. TE jui, A. Gismatulin, V. Volodin, V. Gritsenko, and A. Chin, "All nonmetal resistive random access memory," *Scientific Reports*, vol. 9, p. 6144, 04 2019.

[5] G. Vinuesa, H. García, M. B. González, K. Kalam, M. Zabala, A. Tarre, K. Kukli, A. Tamm, F. Campabadal, J. Jiménez, H. Castán, and S. Dueñas, "Effect of dielectric thickness on resistive switching polarity in tin/ti/hfo_2/pt stacks," *Electronics*, vol. 11, no. 3, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/3/479

[6] R. Guido, "A cmos-compatible bipolar analogue switching redox-based reram with taox/hfo2 bilayer," *Master thesis*, 2021.

[7] T. Stecconi, R. Guido, L. Berchialla, A. La Porta, J. Weiss, Y. Popoff, M. Halter, M. Sousa, F. Horst, D. Dávila, U. Drechsler, R. Dittmann, B. J. Offrein, and V. Bragaglia, "Filamentary taox/hfo2 reram devices for neural networks training with analog in-memory computing," *Advanced Electronic Materials*, vol. n/a, no. n/a, p. 2200448. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/aelm.202200448

[8] *Ginestra User Guide*. Applied MDLx™, 2022. [Online]. Available: http://origin-www.appliedmaterials.com/ko/products/applied-mdlx-ginestra-simulation-software

[9] IV, C. M. U. the B1500A MFCMU, and SCUU.

[10] https://blog.beamex.com/pt100-temperature sensor.

[11] F. Völklein, H. Reith, and A. Meier, "Measuring methods for the investigation of in-plane and cross-plane thermal conductivity of thin films," *physica status solidi (a)*, vol. 210, no. 1, pp. 106–118, 2013.

[12] I. H. Im, S. J. Kim, and H. W. Jang, "Memristive devices for new computing paradigms," *Advanced Intelligent Systems*, vol. 2, 2020.

[13] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," *Proceedings of the IEEE*, vol. 106, no. 2, pp. 260–285, 2018.

[14] H. Y. Peng, L. Pu, J. C. Wu, D. Cha, J. H. Hong, W. N. Lin, Y. Y. Li, J. F. Ding, A. David, K. Li, and T. Wu, "Effects of electrode material and configuration on the characteristics of planar resistive switching devices," *APL Materials*, vol. 1, no. 5, p. 052106, 2013. [Online]. Available: https://doi.org/10.1063/1.4827597

[15] N. Raghavan, "Performance and reliability trade-offs for high- rram," *Microelectronics Reliability*, vol. 54, no. 9, pp. 2253–2257, 2014, sI: ESREF 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0026271414003370

[16] K. Kamiya, M. Yang, B. Magyari-Kope, Y. Nishi, and K. Shiraishi, *Modeling of resistive random access memory (RRAM) switching mechanisms and memory structures*, 12 2014, pp. 262–284e.

[17] T. Sadi and A. Asenov, "Microscopic kmc modeling of oxide rrams," in *Numerical Methods and Applications*, G. Nikolov, N. Kolkovska, and K. Georgiev, Eds. Cham: Springer International Publishing, 2019, pp. 290–297.

[18] Y. Liao, B. Gao, F. Xu, P. Yao, J. Chen, W. Zhang, J. Tang, H. Wu, and H. Qian, "A compact model of analog rram with device and array nonideal effects for neuromorphic systems," *IEEE Transactions on Electron Devices*, vol. 67, no. 4, pp. 1593–1599, 2020.

[19] A. M. Ginestra®.

[20] C. M. M. Rosário, B. Thöner, A. Schönhals, S. Menzel, A. Meledin, N. P. Barradas, E. Alves, J. Mayer, M. Wuttig, R. Waser, N. A. Sobolev, and D. J. Wouters, "Metallic

filamentary conduction in valence change-based resistive switching devices: the case of taox thin film with x 1," *Nanoscale*, vol. 11, pp. 16 978–16 990, 2019. [Online]. Available: http://dx.doi.org/10.1039/C9NR05285B

[21] N. Maity, R. Maity, and S. Baishya, "Voltage and oxide thickness dependent tunneling current density and tunnel resistivity model: Application to high-k material hfo2 based mos devices," *Superlattices and Microstructures*, vol. 111, pp. 628–641, 2017.

[22] K. P. Inc., "Tantalum disc cathodes - thermionic emitters." [Online]. Available: https://www.kimballphysics.com/

[23] X.-b. Lu, K. Maruyama, and H. Ishiwara, "Characterization of hftao films for gate oxide and metal-ferroelectric-insulator-silicon device applications," *Journal of Applied Physics*, vol. 103, no. 4, p. 044105, 2008. [Online]. Available: https://doi.org/10.1063/1.2871772

[24] V. Gritsenko, T. Perevalov, V. Voronkovskii, A. Gismatulin, V. Kruchinin, A. Vladimir, V. Pustovarov, I. Prosvirin, and Y. Roizin, "Charge transport and the nature of traps in oxygen deficient tantalum oxide," *ACS Applied Materials Interfaces*, vol. 10, 01 2018.

[25] J. Robertson and C. W. Chen, "Schottky barrier heights of tantalum oxide, barium strontium titanate, lead titanate, and strontium bismuth tantalate," *Applied Physics Letters*, vol. 74, no. 8, pp. 1168–1170, 1999. [Online]. Available: https://doi.org/10.1063/1.123476

[26] A. Padovani, L. Larcher, O. Pirrotta, L. Vandelli, and G. Bersuker, "Microscopic modeling of hfox rram operations: From forming to switching," *IEEE Transactions on Electron Devices*, vol. 62, no. 6, pp. 1998–2006, 2015.

[27] Z. Zhang, Y. Wu, H.-S. P. Wong, and S. S. Wong, "Nanometer-scale hfo_x rram," *IEEE Electron Device Letters*, vol. 34, no. 8, pp. 1005–1007, 2013.

[28] Y. Lu, B. Gao, F. Xu, J. Tang, H. Qian, and H. Wu, "Real-time-scale 3d kinetic monte carlo simulation for hafnium oxide based rram in 1t1r cell," in *2022 6th IEEE Electron Devices Technology Manufacturing Conference (EDTM)*, 2022, pp. 363–365.

[29] B. Butcher, G. Bersuker, L. Vandelli, A. Padovani, L. Larcher, A. Kalantarian, R. Geer, and D. Gilmer, "Modeling the effects of different forming conditions on rram conductive filament stability," in *2013 5th IEEE International Memory Workshop*, 2013, pp. 52–55.

[30] C. D. Landon, R. H. T. Wilke, M. T. Brumbach, G. L. Brennecka, M. Blea-Kirby, J. F. Ihlefeld, M. J. Marinella, and T. E. Beechem, "Thermal transport in tantalum oxide films for memristive applications," *Applied Physics Letters*, vol. 107, no. 2, p. 023108, 2015. [Online]. Available: https://doi.org/10.1063/1.4926921