Politecnico di Torino, Grenoble INP Phelma, EPFL
IBM Research - Zurich

Master's Thesis

Electrical Characterization of Resistive RAM Devices for Neuromorphic Computing Applications

Supervisors:
Prof. Dr. Carlo Ricciardi
Dr. Antonio La Porta
Dr. Valeria Bragaglia
Dr. Bert Jan Offrein

Candidate:
Luca Berchialla

September 2021

Contents

List of Figures

1 Introduction

In the past years, traditional computing architectures have faced several limitations, e.g. heating, high power consumption and scaling difficulties, especially for artificial intelligence, AI, workloads [44]. In particular the von Neumann architectures, where the computing element is separated from the memory, require transfer of data from the memory element to the computing element and vice versa. With the increase of computational power, larger amounts of data are fed into the processor. However, while computing has made huge leaps forward in terms of raw computational power and energy efficiency, memory and the interconnections between computing and memory elements have failed to keep up. This phenomenon is called the memory wall [41]. For many workloads, the gap between memory power and computing power means that more time and energy is spent on moving the data than on computation itself [16].

For this reason, in-memory computing is one of the techniques being developed to overcome these limitations. In-memory computing radically changes the von Neumann architecture because the calculations on the data and the data storage are in the same location [18]. Naturally, this approach is less flexible than the von Neumann architectures, but it is particularly useful for accelerating inference and training of artificial neural networks, ANNs.

Deep neural networks, DNNs, are a subset of ANNs with more hidden layers, Figure 1.1. In ANNs and DNNs the neuron layers can be represented by vectors and the synaptic weights by matrix elements. The output of the first hidden layer is computed by multiplying the input layer vector with the matrix of the synaptic weights associated to the input layer. This procedure is called vector matrix multiplication, VMM, and it is repeated for all the following layers up to the output layer to generate the ANN or DNN output. During training, an enormous set of input and output vectors is provided to find the synaptic weights that can produce output vectors with minimum deviation from the provided ones. For training, various training algorithms exist such as the stochastic gradient descent, SGD, algorithm. This algorithm approaches the optimum parameters iteratively by adding a calculated value to the current parameters, such that the output of the network takes more resemblance to the desired output. During inference, the input vector is provided as input and the output vector can be computed since the synaptic weights are known from training.

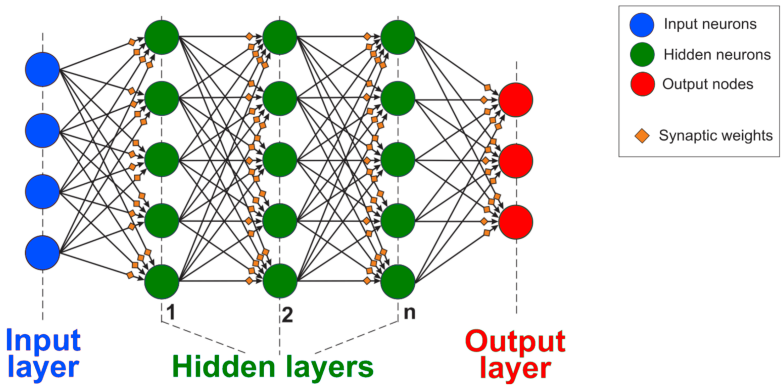

Figure 1.1: Scheme of a fully connected deep neural network, DNN, with the input layer, the hidden layers and the output layer highlighted by different colors.

Neuromorphic computing is a field in which the computational paradigms are inspired by the brain. One branch of neuromorphic computing is focused on accelerating ANN and DNN workloads by making use of memristors in a crossbar structure to perform VMM. In Figure 1.2, such a crossbar array is depicted. The crossbar array is a structure composed of two layers of perpendicular metal lines. One line is used for input, the other for output. A memristive element connects the two

lines at each intersection. The memristor is an ideal circuit element which "changes its resistance depending on how much charge flows through it" [20, 10], meaning it behaves like a resistor with memory. This property is used program and store a synaptic weight in each intersection of the crossbar array. In doing so, VMM is possible.

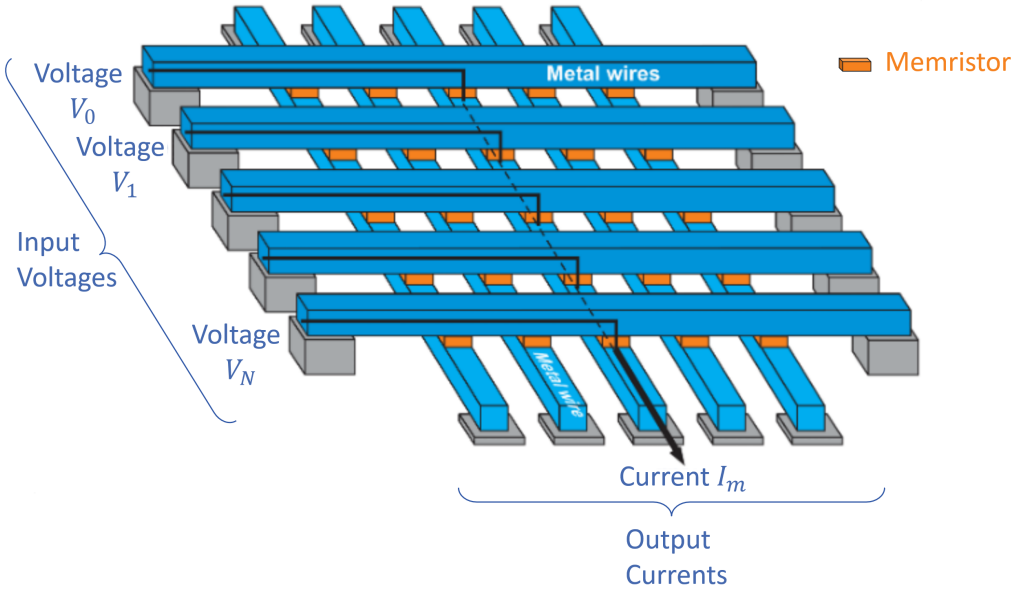

Figure 1.2: Schematic of a crossbar array. The inputs are on the left, the outputs are at the bottom and the memristive elements are colored orange.

Assuming the metal lines have a negligible resistance with respect to the memristive elements, it is possible to compute the current for each output metal line, I_m, when given the voltage inputs, V_i, and the conductance of each memristor, G_{mi}, as in Equation 1.1 [18].

$$I_m = \sum_{N}^{i=0} G_{mi} \cdot V_I \tag{1.1}$$

This equation expresses a summation of the output vectors. The total output current results from the multiplication between the vector of input voltages, $\vec{V}$, and the matrix expressing the conductance of the memristive elements, G. This is the physical implementation of vector matrix multiplication in crossbar arrays, where the inputs are given as voltages and the coefficients of the matrix are the conductance values of the memristors. Since memristors are memory elements, the computation happens in the same place as where the information is stored. This is the implementation of an in memory computing architecture.

2 Theoretical background

2.1 Resistive RAMs

To implement memristors, different technologies are used, such as: resistive RAM, RRAM; phase change memory, PCM; magnetoresistive RAM, MRAM; ferroelectric RAM, FeRAM [18]. In this thesis RRAM devices are analyzed.

RRAMs are generally metal-insulator-metal, MIM, structures. The resistive switching is caused by redox reactions between the insulator and one of the metal electrodes. Among the used insulator materials are numerous metal oxides such as hafnium oxide HfO_x, titanium oxide TiO_x, tantalum oxide TaO_x, nickel oxide Ni_O, manganese oxide MnO_x and aluminum oxide AlO_x [43]. For the electrodes, the commonly used materials are metals such as aluminum Al, titanium Ti, copper Cu, tungsten W, platinum Pt, titanium nitride TiN and tantalum nitride TaN [43].

In this thesis RRAMs based on two different material stacks are analyzed and compared. One material stack is used as a baseline since it is a common structure in literature. The baseline is composed of $TiN - HfO_x - Ti - TiN$, Figure 2.1 a. The second material stack is similar to the first, but uses a sub-stoichiometric transition metal oxide, MO_x, to replace the Ti layer, Figure 2.1 b.

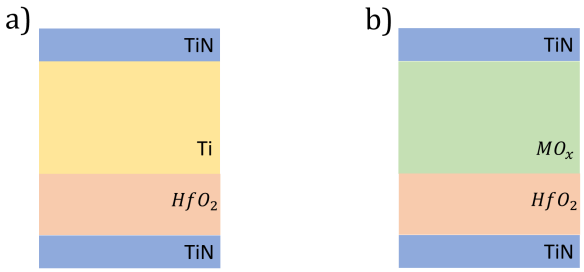

Figure 2.1: Schematic of the 2 analyzed RRAM material stacks: a) $HfO_2 - Ti$ technology and b) $HfO_2 - MO_x$ technology.

After device fabrication, RRAMs behave like capacitors since there is a strongly insulating material between the two electrodes. In order to enable the resistive switching, RRAM devices should undergo a process known as electroforming or forming. This process is performed by applying a voltage between the electrodes high enough to cause a partial breakdown of the insulating layer: a conductive filament, i.e. a low resistance path of defects, is created in the dielectric, Figure 2.3 c [31]. In $HfO_2 - Ti$ devices, oxygen vacancies/ion pairs are generated during forming. Due to the fact that a positive voltage is applied to the top electrode, oxygen ions tend to flow towards the TiO_x layer which is formed at the interface with the HfO_2. The oxygen ion flow leaves a conductive path through the HfO_2, consisting of oxygen vacancies.

Subsequently, RRAM devices are able to show resistive switching upon the application of bipolar voltage sweeps or a sequence of pulses. For $HfO_2 - Ti$ and most RRAM devices it is possible to increase the resistance of the device by applying a negative voltage, which transitions the device to the high resistance state, HRS [17]. This operation is called Reset. The reset operation pushes the oxygen ions from the TiO_x layer back into the HfO_2 layer, resulting in recombination with the oxygen vacancies in that layer. Recomination weakens the conductive filament until its interruption, Figure 2.3 d [31].

Instead, for $HfO_2 - Ti$ devices, the Set operation is performed by applying a positive voltage, which lowers the resistance of the device and transitions it to the low resitive state, LRS[17]. A positive voltage, in analogy to forming, facilitates the generation of oxygen vacancies/ions pairs,

thus re-establishing the conductive filament and causing the oxygen ions to drift towards the top electrode [31].

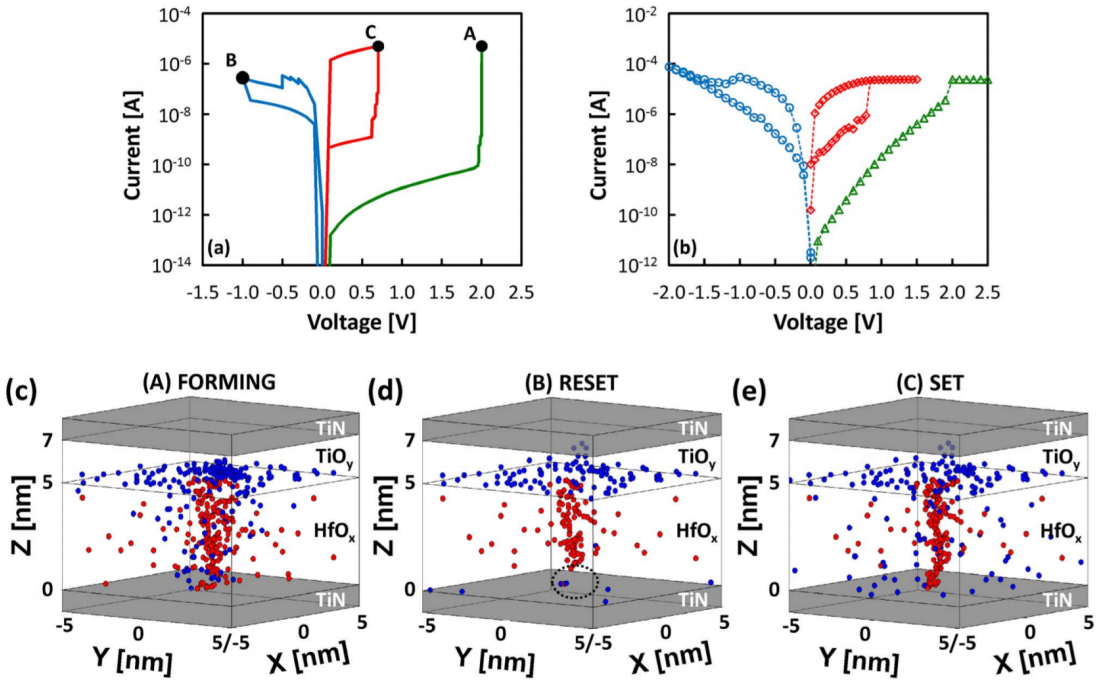

Figure 2.2: "Simulation of a forming-RESET–SET cycle with ICC = 5 μA. I–V characteristics (a) simulated and (b) measured during a full forming-RESET–SET cycle (a 25-μA current compliance is imposed). 3-D oxygen vacancy (red spheres) and ion (blue spheres) distributions obtained from simulations at the end of (c) forming, (d) RESET, and (e) SET operations."[31]

2.2 Resistive processing unit

To successfully accelerate DNN inference and training, it is necessary to build a resistive processing unit RPU, composed of multiple crossbar arrays for full implementation of the structure of DNNs. As described before, the synaptic weights are mapped to the conductance range of the RRAMs in the crossbar arrays. In the ideal case, the conductance of RRAM devices would change linearly between the LRS and the HRS upon the application of bipolar pulses. Pulses that produce an increase in conductance are called potentiation pulses; instead when the pulses cause a decrease in conductance they are called depression pulses.

During training the synaptic weights of the crossbar array are changed by applying the required number of potentiation or depression pulses to the devices, depending on the given output of the SGD algorithm. However, only in ideal devices the conductance changes linearly with the number of applied pulses. To compensate for asymmetries and non-linearity a novel training algorithm, designed especially for RRAMs, has been developed by IBM [21, 12]. This algorithm is called Tiki-Taka and it exploits a characteristic of RRAM devices: the symmetry point, SP, which is the conductance value at which a single potentiation or depression pulse causes the same absolute conductance change. The synaptic weights can be mapped in order to have the SP at weight = 0. For training with the Tiki-Taka algorithm two identical crossbar arrays are needed. One crossbar array has all the devices set at the symmetry point and it acts as a reference for the second one, which instead is updated for training of the neural network. It has been proven that the Tiki-Taka algorithm can

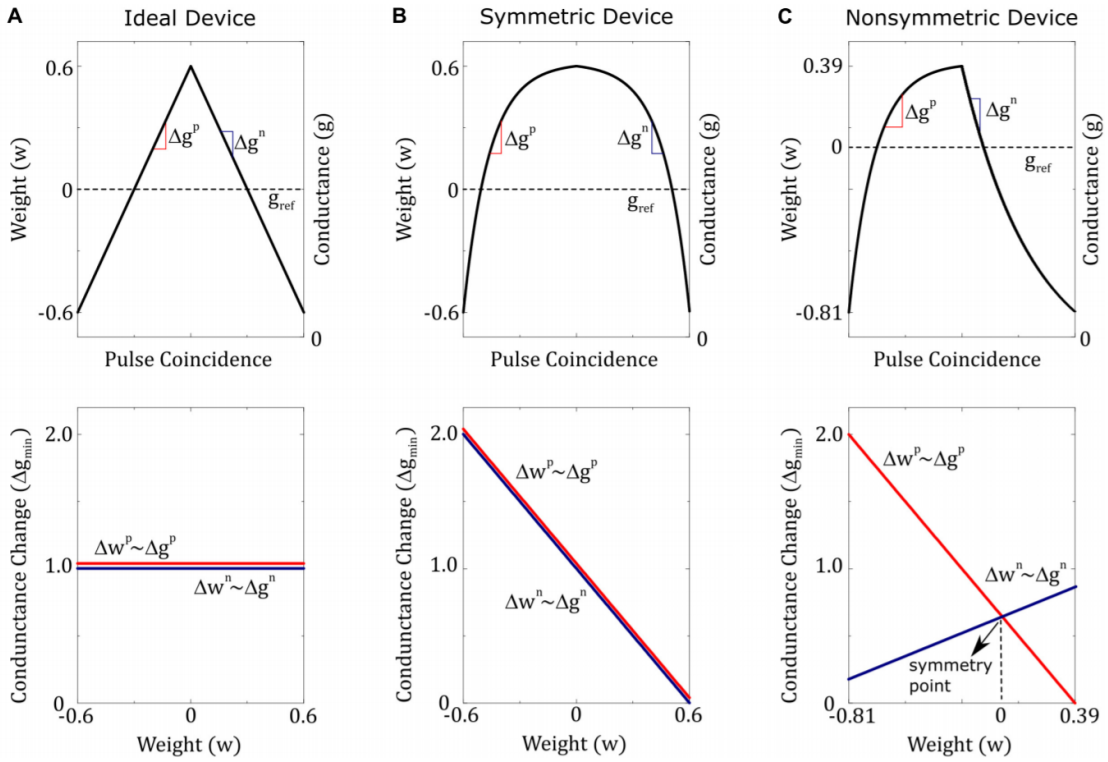

Figure 2.3: "Three different device switching characteristics are illustrated. (A) Ideal device: conductance increments and decrements are equal in size and do not depend on device conductance. (B) Symmetric device: conductance increments and decrements are equal in strength, but both have a dependence on device conductance. (C) Non-symmetric device: conductance increments and decrements are not equal in strength and both have different dependencies on device conductance. However, there exists a single point (conductance value) at which the strengths of the conductance increment and decrement are equal. This point is called the symmetry point and for the illustrated example matches the reference device conductance and hence happens at w = 0. "[12]

significantly improve the performance accuracy of DNNs trained on crossbar arrays with RRAMs [21, 12].

3 Methodology

In this section, the methods used in this thesis for investigating the behavior of memristive devices are described. In general, a quasi-static (DC) and pulsed (AC) characterization are carried out for each device. Moreover, explanations of additional experiments and data analysis techniques are given in this section.

3.1 Quasi-static (DC) Electrical characterization

The Quasi-static (DC) electrical characterization is performed by applying a slowly varying signal to an RRAM device. This methodology neglects transient effects and assumes that the system is electrically in a steady state.
In the reported measurements, this condition is approximated by applying a voltage staircase ramp. For the sake of device-to-device comparison, the hold time and voltage step are fixed: $t_{hold} = 0.1\,s$ and $V_{step} = 20\,mV$. Figure 3.1 shows the voltage staircase ramp as applied for the DC electrical characterization.

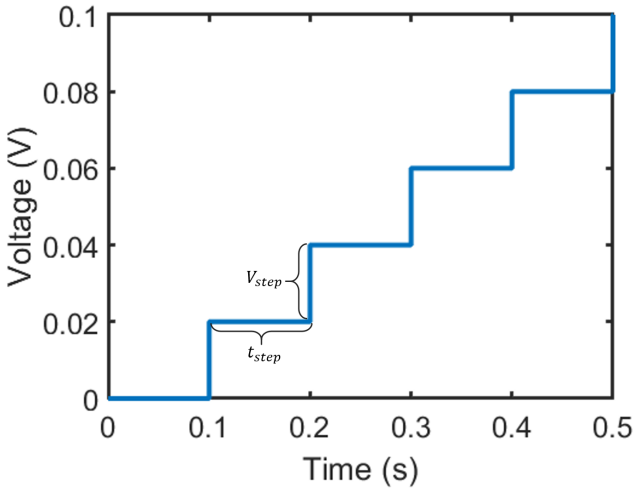

Figure 3.1: Example of a voltage staircase ramp with $t_{hold} = 0.1\,s$ and $V_{step} = 20\,mV$.

3.1.1 Setup

The measurements are taken from a Süss Microtec probe-station, Figure 3.2, equipped with an Agilent 4115C semiconductor parameter analyzer.
The measurements are carried out with multiple Source Measurement Units (SMUs) of the analyzer. Figure 3.3 shows a simplified electrical circuit schematic of an SMU. Each SMU can force voltage or current and contemporarily measure voltage and current.
The probe-station provides electromagnetic shielding through its closed metal chassis. Chips with RRAM devices are held in place by a vacuum hose system on the gold plated ceramic chuck. The chuck realizes the electric contact to the chip substrate, which is in direct contact with the bottom electrode of the RRAM devices.
The top electrode pads are contacted by means of electrical probe tips mounted on Cascade Microtech DPP210 3 axis precision positioner systems. The connection between the semiconductor parameter analyzer and the probe-station is made with triax cables, Figure 3.4.
Triax cables have three concentric conduction paths. The signal is conducted by the central connector and it is protected by the guard, which is forced to the same voltage as the signal by a buffer internal to the SMU under use. The outer connector, the shield, is instead forced at the circuit common

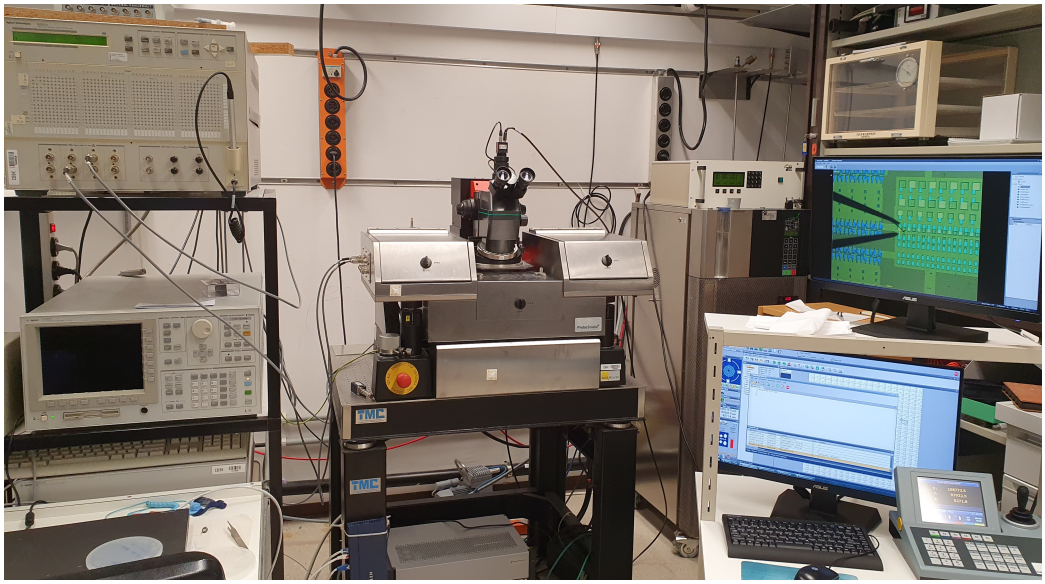

Figure 3.2: Picture of the DC characterization probe-station.

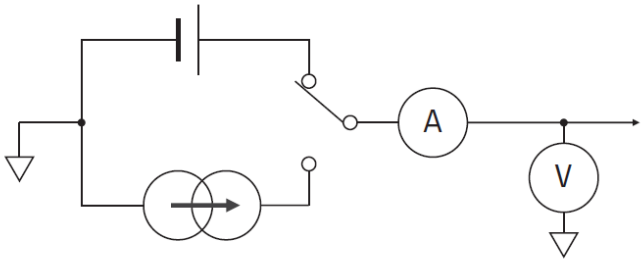

Figure 3.3: Simplified schematic of a Source Measurement Unit (SMU), from [37].

voltage. In the described set-up, the circuit common voltage is tied to the earth ground. In standard coaxial cables the guard is not present, thus the conductor surrounding the signal is the shield. Triax cables enable a better capacitive decoupling of the signal, since the signal and the guard are forced at the same voltage and therefore a smaller leakage current will flow to the signal during a transient.
For a triax cable, supposing that the buffer forces the guard voltage V_g within one part per million of the signal [37], the leakage current can be calculated as:

$$i_{leak} = \frac{V_g - V_s}{R} = \frac{1.000\,001\,\mathrm{V} - 1\,\mathrm{V}}{1 \times 10^9\,\Omega} = \frac{1 \times 10^{-6}\,\mathrm{V}}{1 \times 10^9\,\Omega} = 1 \times 10^{-15}\,\mathrm{A} = 1\,\mathrm{fA} \tag{3.1}$$

An estimation of the leakage current in a coaxial cable can be calculated assuming that the insulation material has a resistance $R = 1\,G\Omega$, an applied signal $V_s = 1\,V$ and the shield connected to ground [37]. This leads to a leakage current:

$$i_{leak} = \frac{V_s}{R} = \frac{1\,\mathrm{V}}{1 \times 10^9\,\Omega} = 1 \times 10^{-9}\,\mathrm{A} = 1\,\mathrm{nA} \tag{3.2}$$

This leakage current is too high for taking measurements in the pA (1×10^{-12} A) range.
The leakage current using a triax cable is dramatically lower with respect to the coaxial cable and its absolute value allows measurements in the pA range without interfering.

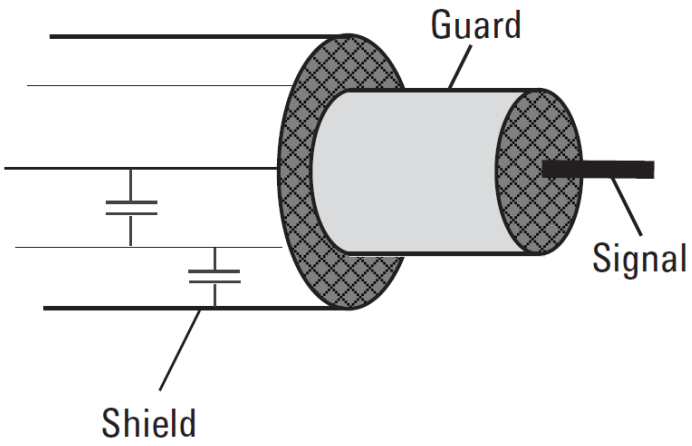

Figure 3.4: Schematic of a triax cable, from [37].

Automated chuck movement During the last weeks of the thesis, the controller of the chuck motors was fixed. Therefore the probe-station could be programmed to move the chuck, and thus the chip, in X,Y,Z. This feature enables automated repeated measurements on different devices.
A custom Labview interface, used to control the chuck movement of another probe-station, was ported and adapted for this probe-station.
The interface runs on a regular computer and communicates to the probe-station trough the General Purpose Interface Bus (GPIB). To perform automatic measurements, a map of the chip must be supplied. The map contains the X,Y coordinates of all the top electrode pads on the chip and for each pad it contains relevant information about the device. By specifying an initial position of the tips with respect to the chip it is possible to establish a bi-univocal relation between the chuck X,Y space and the map X,Y space. This allows the user to choose a device and move the chuck to bring the selected device below the tips.

Custom measurement interface The semiconductor parameter analyzer is remote controlled wit a custom Labview interface, as shown in Figure 3.5. The interface runs on a regular computer and controls the tool using the GPIB bus. The custom Labview interface permits extreme flexibility for the measurements.
The program allows the user to specify the SMU with which a voltage or current sweep should be performed and the SMU which will remain tied to 0 V. This function has been implemented to avoid manual connection and disconnection of the triax cable to the desired probe tip. The manual connection was introducing undesired charges in the system, which are potentially harmful for the device under test, DUT. In order to avoid partial breakdown on the devices before starting a measurement, a fixed procedure should be followed: a passive area on the chip should be contacted and the program should be started and executed without any set measurements. By doing so, the tool forces 0 V on all active SMUs and therefore eliminates any potential accumulated charge. After this step it is safe to contact a device.
In each column of the "step input", Figure 3.5, a voltage or current sweep can be defined. For each column the settings of the sweeps can be set independently.
The compliance sets the maximum current (or voltage) that the tool will allow when performing a voltage (or current) sweep.
For example, if we would perform a voltage sweep between 0 V and 2 V with a compliance current of 1 mA on a 1 kΩ resistor, we would arrive at $V_{applied} = 1\,V$ and $I_{measured} = 1\,mA$. At this point the tool will try to force a voltage $V_{applied} > 1\,V$ and therefore it needs to flow a current $I > 1\,mA$. However, the compliance current limits the current that can flow to 1 mA and therefore the tool will force a voltage as close as possible to the targeted voltage while maintaining the applied current below or equal to the compliance current. In this example 1 V.

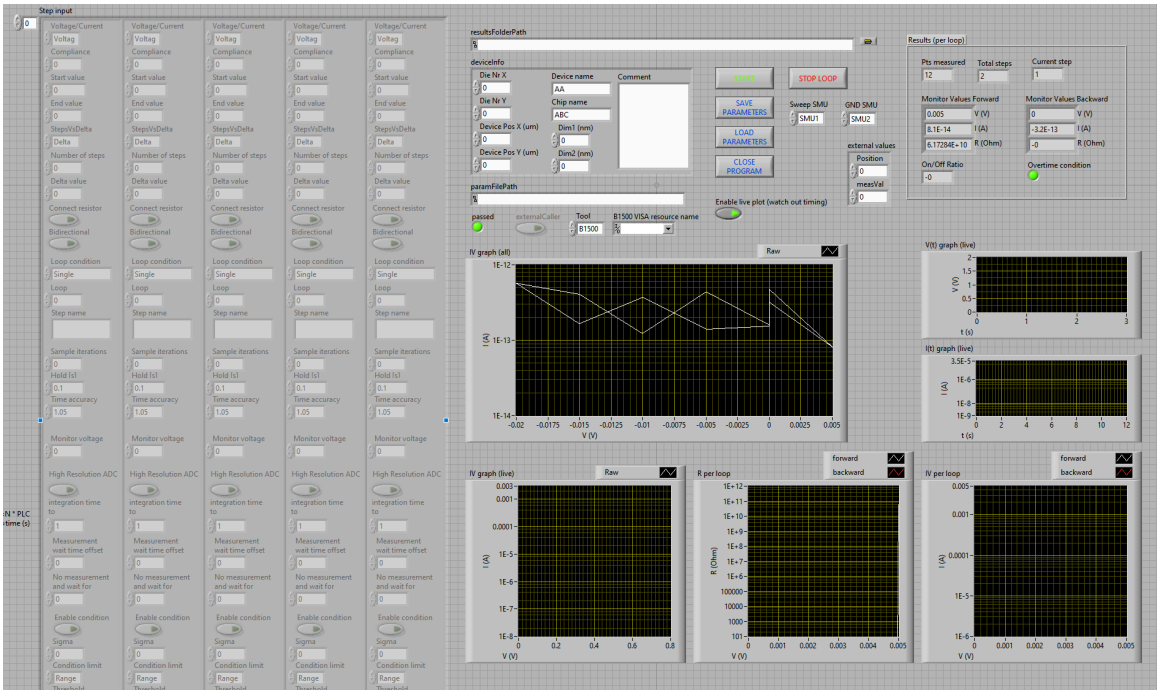

Figure 3.5: Screen capture of the custom Labview interface for DC electrical characterization.

It is also possible to set the start and end values of the sweep and to set the number of steps in which the staircase sweep will go from the start to end value, or alternatively the step size. If the bidirectional switch is enabled, the sweep will be performed from start value to end value and subsequently from end value to start value; otherwise the sweep is only from start value to end value. Other parameters of the staircase sweep can also be set, such as Wait time t_{wait} and Hold time t_{hold}, Figure 3.6.

The hold time is the amount of time that the SMU forces a certain value before going to the following step of the staircase sweep. The hold time should be large enough to ensure that the SMU output is stable before going to the next step.
The wait time is the amount of time that the SMU forces a certain value before taking a measurement. The wait time should be large enough to ensure that the SMU output is stable before performing the measurement.
In addition, it is possible to set the number of measurements performed for each step of the staircase sweep and the integration time of the ADC of the SMU.
The user can specify whether a specific column or multiple columns in the interface should be enabled, disabled or repeated a certain amount of times. Using these functions it is possible to apply multiple sweeps to the DUT, which can be executed iteratively and/or looped.
The interface also shows a live IV plot and gives the user the possibility to manually interrupt the sweep at the current value of the voltage or current by pressing the stop button.

3.1.2 Pristine Sweeps

Pristine sweeps are voltage sweeps performed on a device in pristine state, where neither partial or full breakdown is present. Typically, a pristine device is in a very high resistance state ($\geq$GΩ) due to the MIM layer stack. The end voltage for these sweeps must be chosen below the forming voltage, since upon electroforming the device is in a lower resistance state than the initial pristine state.

The pristine sweep gives the leakage current through the dielectric stack of the device, therefore it can be considered as a method to electrically characterize the material stack after fabrication.

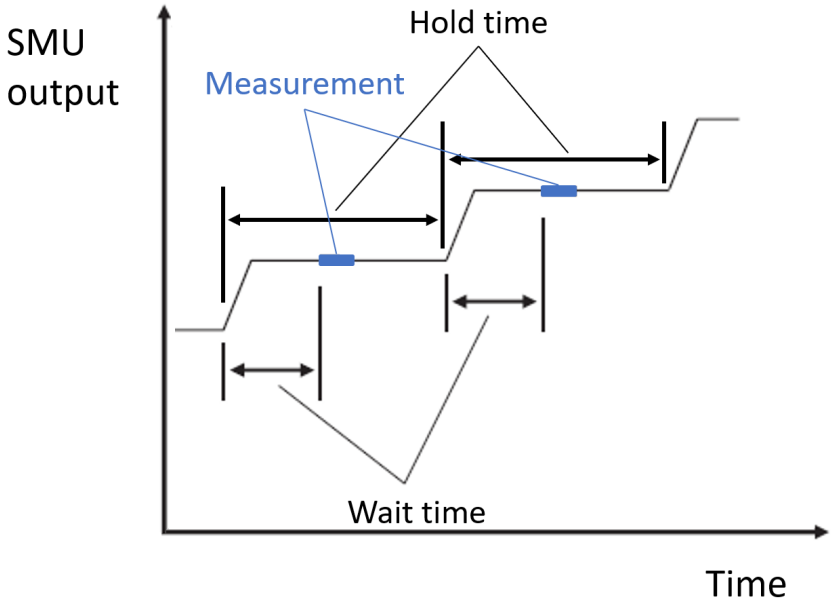

Figure 3.6: Staircase sweep Wait time t_{wait} and Hold time t_{hold} detail.

Since the pristine sweeps characterize the material stack after processing, they can give information about the uniformity of the used fabrication processes wihin a chip between different chips.

During routine measurements, pristine sweeps are useful to detect non-functioning devices. These devices are either shorted due to processing problems or have been subject to unintentional partial or full breakdown. The listed conditions are easily detectable because the devices in pristine state are very restive $R >> 100\,M\Omega$, while non-functioning devices have low resistances $R << 10\,k\Omega$.

3.1.3 Forming

The objective of forming is to induce a partial, or soft, breakdown in the dielectric material stack. Historically, the breakdown of oxides has been of great interest in assessing the reliability of gate oxide in CMOS technology [26]. Specifically, this reliability is studied as the time-dependent dielectric breakdown, TDDB, which is the required time for a dielectric to breakdown under a constant voltage stress, CVS. In this experiment the breakdown happens when the current dramatically increases, since the oxide loses its insulating properties.
Other experiments on thin oxides can be carried out to precisely characterize the breakdown processes. In constant current stress, CSS, experiments a constant current is applied to the dielectric and the breakdown is detected when the measured voltage drops suddenly [36].
The TDDB can be related to other experiments such as the voltage-ramp dielectric breakdown, VRDB, or the current-ramp dielectric breakdown, IRDB, [2]. In VRDB the breakdown voltage V_{bd} is the voltage at which a sudden increase of current is observed. In IRDB the V_{bd} is the highest measured voltage in the measurement, Figure 3.7.
The VRDB and IRDB are much quicker tests than the CVS or the CCS if they are performed at meaningful voltages or currents, since the breakdown is more likely to happen by raising the voltage rather than waiting more time [40]. In literature on RRAM characterization, VRDB is commonly called voltage forming and IRDB is commonly called current forming [17].
Since the goal of forming is to induce a soft breakdown, during voltage forming a compliance current must be set to control the maximum current that can flow through the DUT. The compliance current

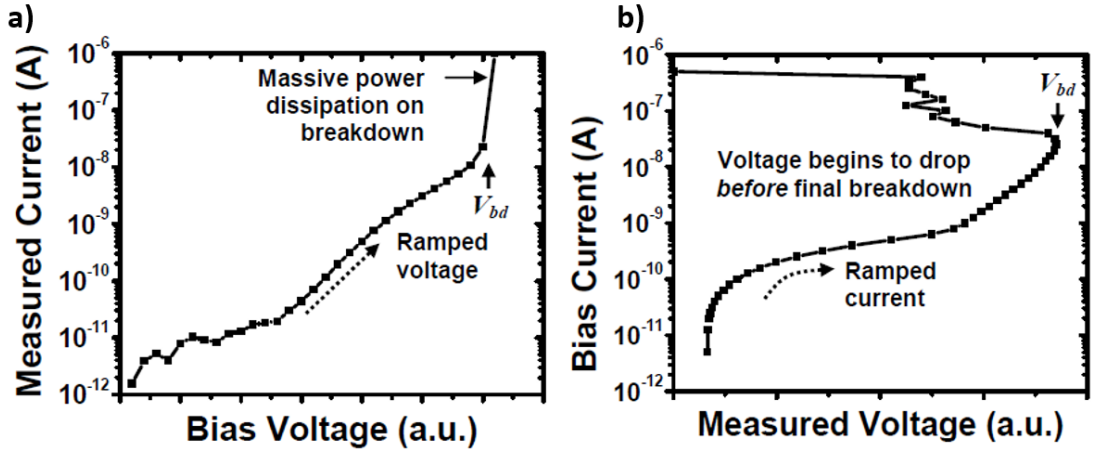

Figure 3.7: Typical VRDB (a) and IRDB (b) current-voltage (I-V) curve showinng the breakdown voltage V_{bd}. Adapted from [2].

prevents the measurement tool from providing too much current to the device. Upon the onset of the soft breakdown, the resistance of the insulating stack is reduced by several orders of magnitude. A large current flow will lead to a hard breakdown, i.e., to am irreversible low resistance ($<500\,\Omega$) state of the device, hence a compliance current is used. Current forming can intrinsically solve this problem since the current is raised by a fixed amount for each step. When the device resistance is reduced by the soft breakdown the measured voltage drops. Thus, with this approach there is no direct danger of providing too much current to the device.

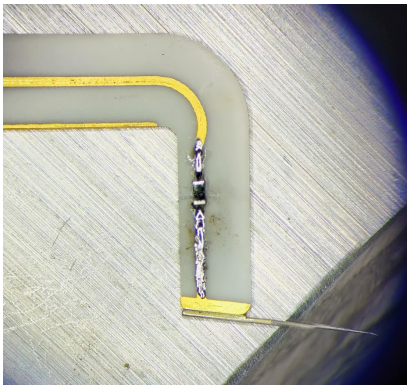

Figure 3.8: Picture of a probe tip with a soldered SMD series resistor.

The measurement tool has a feedback loop to force the desired voltage or current with given compliance values. Although the compliance feedback loop mechanism in the SMU is accurate, its time response is orders of magnitude slower with respect to the time scale of the breakdown process. Therefore, for a certain amount of time, which can be quantified in the ms range, the voltage over and the current through the device are unknown. To mitigate this effect, an SMD series resistor R_s is soldered in series to the signal line of the probe tip, Figure 3.8. This resistor limits the voltage drop across the DUT during the transient and reduces the chances of a hard breakdown.

The forming voltage V_f is defined as the breakdown voltage V_{bd} and the forming current I_f is defined

as the current flowing at V_f, just before the breakdown, as shown in Figure 3.9.

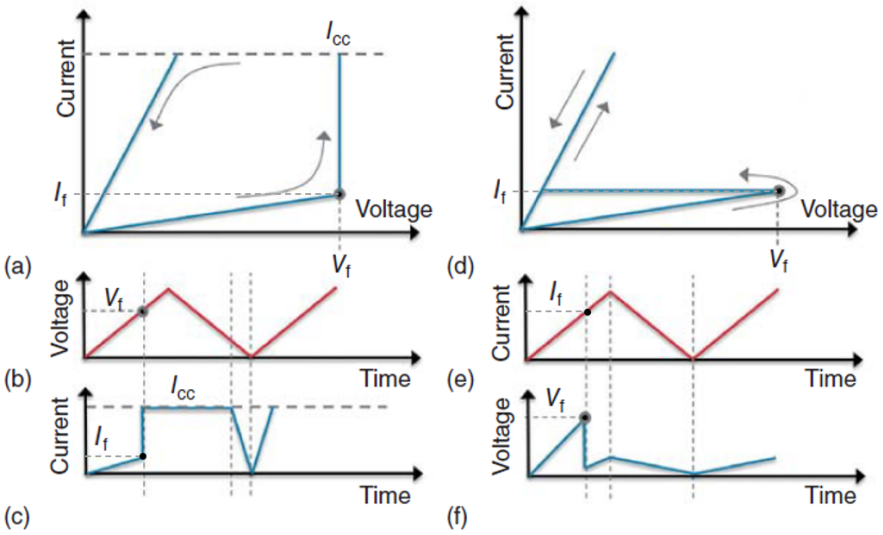

Figure 3.9: "Schematics of the current–voltage behaviors undergoing electroforming with applied (a) voltage sweeps and (d) current sweeps. The applied input sweeps in a time domain for each case are illustrated in (b) and (e), respectively. The corresponding output current and voltage in the time domains are also illustrated in (c) and (f), respectively" [17]. Adapted from [17].

3.1.4 Cycling

After forming, or soft breakdown, the devices can be cycled using bipolar voltage sweeps between the HRS and the LRS. During cycling, the resistance of the HRS and LRS is read at $\pm 0.2\,\mathrm{V}$. With the HRS and LRS values known, it is possible to calculate the On-Off ratio defined as HRS/LRS, also termed memory window. Other figures of merit such as the cycle-to-cycle variability and the abruptness of the transition can be calculated. The cycle-to-cycle variability is evaluated by calculating the standard deviation of the HRS and/or LRS values of all the performed cycles. The abruptness of the HRS/LRS transitions can be expressed by calculating the maximum change in resistance within 100 mV during the Set or Reset transition and normalizing it to the HRS value.

The values of the HRS and LRS, the stability upon repeated cycling and the shape of the IV curve are not only intrinsic properties of the material stack, but they also depend on the parameters of the applied sweeps and on the history of previous measurement sweeps. This highlights the need for developing a consistent and repeatable measurement procedure that allows the comparison of different devices and the extraction of relevant statistical information. The measurement protocol is optimized to maximize the performance of the devices while maintaining a low device-to-device variability, but it should also resemble the way that the device will be operated in its final applications.

HfO_2 – Ti devices After forming, the HfO_2 – Ti devices have a resistance close to the LRS and a reset to the HRS should be performed. However, this reset transition is different from the subsequent ones because it can happen in a large voltage range and the voltage sweep has to be stopped with $0.1 - 0.2\,\mathrm{V}$ after the reset transition. Stopping the voltage sweep avoids a complementary set transition to a hard breakdown, i.e. a permanent filament condition.
After the first reset the device can be cycled between LRS and HRS by applying bipolar voltage sweeps from 0 V to V_{stop_set} to 0 V and then from 0 V to V_{stop_reset} to 0 V. The V_{stop_set} should be

higher by $0.2 - 0.5\,\mathrm{V}$ than the actual set transition voltage to give room for cycle-to-cycle variability and device-to-device variability. The V_{stop_reset} should not exceed $0.1 - 0.2\,\mathrm{V}$ above the reset transition voltage to avoid a complementary set to a permanent filament state, similar to the first reset.

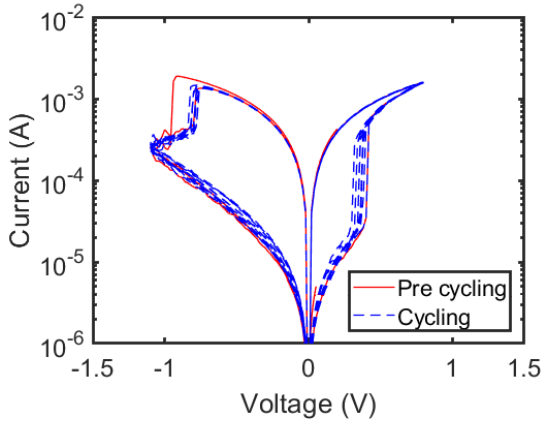

Figure 3.10: Current-voltage, I-V, curve of a $HfO_2 - Ti$ device with the pre-cycling sweeps in solid red and the repeated cycling in dashed blue.

$HfO_2 - MO_x$ devices After forming, the $HfO_2 - MO_x$ devices have a resistance higher than the $HfO_2 - Ti$ devices and they do not reset further, thus a set transitio should be performed. The device can be directly cycled with constant V_{stops}.Typically, at this point large cycle-to cycle-variability, as well as large device-to device-variability and small endurance in cycling is obtained. Therefore, a stabilization procedure, also known as "wake-up" procedure, has been developed. This procedure consists of a first set at relatively high voltages $1.8 - 2\,\mathrm{V}$, using a compliance current to prevent the device from reaching a too low resistance state. Afterwards, the device is cycled with decreasing V_{stops} until the chosen V_{stops} for repeated cycling and device characterization.
The V_{stops} are chosen to maximize the HRS/LRS ratio, thus they correspond to the maximum V_{stops} which produce cycling with little or no complementary switching. The choice of the V_{stops} can be made for each device, for each device size or for each technology.

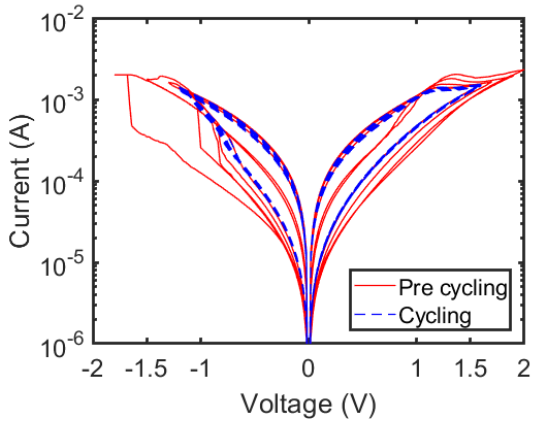

Figure 3.11: Current-voltage, I-V, curve of a $HfO_2 - MO_x$ device with the pr-cycling sweeps in solid red and the repeated cycling in dashed blue.

3.2 Pulsed (AC) Electrical characterization

Training and inference of ANN based on SGD or SGD-like update algorithms are typically implemented via short applied electrical pulses to a set of devices interconnected in one or multiple arrays of matrices. Therefore, to mimic real operations on such architectures, a rather comprehensive pulsed characterization (AC) of the RRAM devices is required.

In AC characterization, the device is stimulated with square voltage pules. As a result, the device changes its resistance by Δ_R, which in general depends on the pulse amplitude, the pulse width and the current device resistance state. For bipolar memristors, the polarity of the voltage pulse determines if the conductance increases, termed potentiation, or decreases, termed depression. In an RPU, the weights are mapped to the conductance (or resistance) state of the single memristive element, thus the resistance change upon the application of a pulse is also termed weight update.

During ANN training the SGD algorithm is generally used [13]. The SGD algorithm approaches

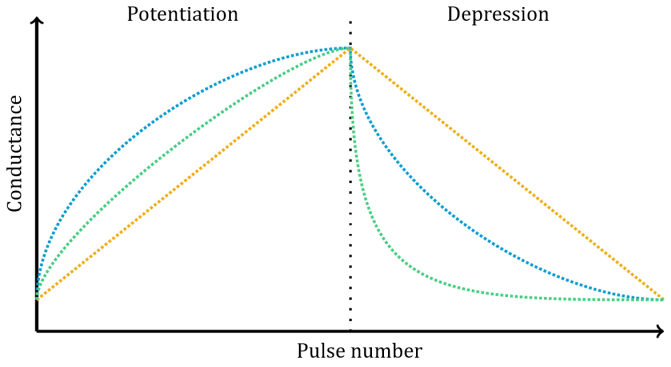

Figure 3.12: Ideal potentiation and depression curves as function of the number of applied pulses. In yellow: linear and symmetric; in blue: non-linear and symmetric; in green: non-linear and asymmetric behaviour.

the optimum parameters by iteratively updating the weights in a new set of calculated values at each training step, i.e. epoch. For SGD-like DNN weight update, the ideal memristive characteristics are linear and symmetric weight update. Intuitively, this guarantees the maximum reproducible weight update over the various training epochs, eliminates the need for weight verification feedback loops and minimizes the reading errors in the inference phase. More specifically, if the weight update is linear, the resistance change at each pulse does not depend on the current resistance, thus it is always equal. If the weight update is symmetric, the device reacts in the same way for potentiation and depression. Having a memristive element that can change its resistance by a fixed amount at each pulse, regardless of the current resistance state, results in a straightforward implementation of the SGD in a crossbar array.
However, real memristive devices can be non-linear and asymmetric, but it has been shown that a moderate non-linearity in the memristor characteristics can be used to introduce regularization, a common method in DNN training to reduce overfitting and improve the performance [21].

3.2.1 Setup

The AC measurement set-up is composed of a 3D printer, WANHAO Duplicator i3 mini, with a custom-built copper sample stage, as depicted in Figure 3.13. Instead of an extruder, a probe tip holder with a camera is installed in the set-up to contact the devices. The top electrode of a device is contacted by the probe tips and the back contact is made through the custom copper chuck.
The probe tip holder can move in X and Z, while the sample stage moves in Y. The movements are controlled with a custom Labview program. The measurement are performed by a National

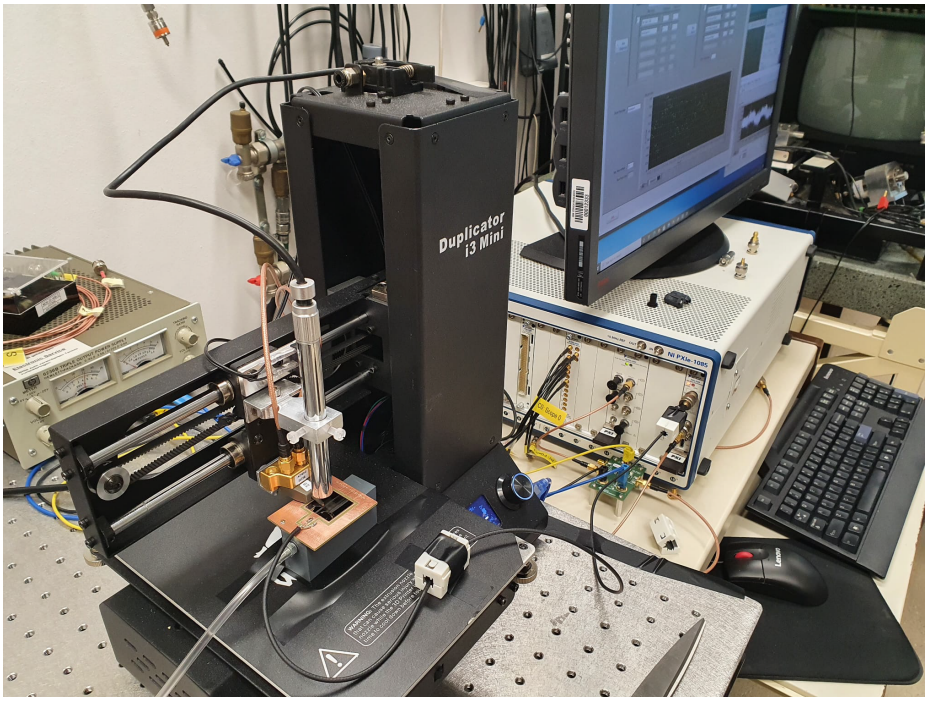

Figure 3.13: Picture of the AC characterization setup.

Instrument $PXIe-1085$ equipped with an $NI\ PXIe-5451$ arbitrary waveform generator and an $NI\ PXIe-5164$ oscilloscope.

The devices are contacted at the top electrode by the probe tips. The left tip is connected to the arbitrary waveform generator by a coaxial cable, which supplies the signal to the device. The right tip is connected to the oscilloscope by a coaxial cable and it serves as a $50\,\Omega$ voltage divider which attenuates transient effects and helps to verify the quality of the contact. The bottom electrode of the device is connected to the chuck, which is connected by a coaxial cable to the oscilloscope. The electric circuit schematic of the set-up is reported in Figure 3.14.

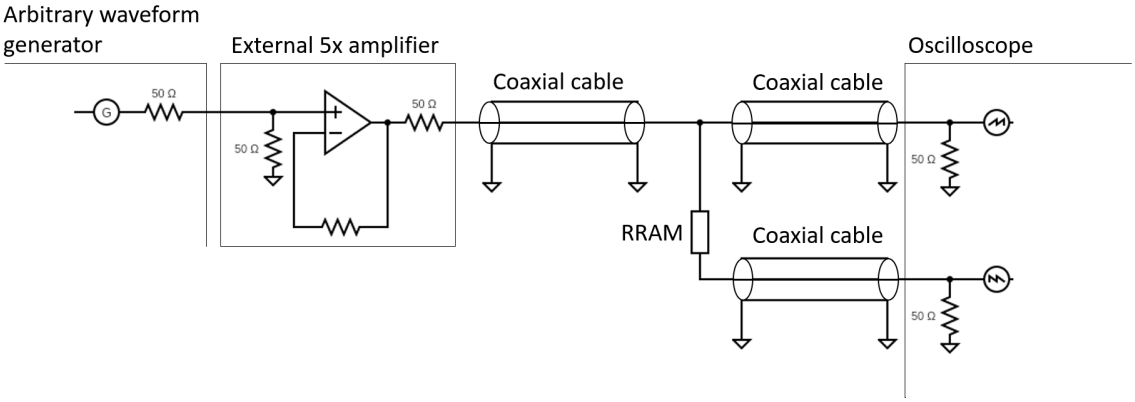

Figure 3.14: Schematic of the AC characterization setup.

The oscilloscope and the arbitrary waveform generator are controlled by a custom Labview program. Within this program it is possible to choose the pulse amplitude, pulse width, slew rate, and the amount of pulses that will be sent to the device.

There are two main kinds of pulses which are sent to the device: reading pulses and programming pulses. To measure the device response after each programming pulse, the device resistance is measured with a reading sequence.
Reading pulses are very long and they are at the voltage V_{read}. For the measurements reported in this thesis the reading pulses have the following parameters: width 8 µs and amplitude $V_{read} = 0.2\,V$. The reading sequence is composed of four consecutive reading pulses; they are at V_{read}, $-V_{read}$, V_{read} and $-V_{read}$. The resistance value is read for all the four pulses and averaged.
Programming pulses have amplitudes of $V_{potentiation}$ or $V_{depression}$, while the pulse width and slew rate are tuned according to the device capacitance, thus the device size.

3.2.2 Choice of the parameters and testing procedure

The devices tested in AC are chosen among the best discovered by the DC characterization; in particular, stability and sufficiently large $On - Off$ ratio are the decisive characteristics.
The width of the programming pulses is initially set to a large value, usually 1 µs to allow the neglection of unwanted transient effects.
However, short pulses are desirable since they decrease the programming time. The slew rate of the signal is set as fast as the instrument is able to drive the specified pulse on the device. Since the devices can be assimilated to leaky capacitors, the larger the area of the device is, the larger its capacitance is. Therefore, big devices are more capacitive, thus harder to drive. This poses a limitation on the overall pulse length of the exciting signals. In fact, a large capacitance needs a smaller slew rate of the signal which limits the minimum time width of a pulse.

After the choice of the slew rate and the pulse length, the device is stimulated with 200 potentiation or depression pulses, depending on the device state, with increasing $V_{potentiation}$ and $V_{depression}$ until a clear response is measured. The starting point for the pulse amplitudes is 0.3 V smaller than the DC V_{stop_set} and V_{stop_reset}.
Once the device responds to the pulses, the same procedure is applied for the opposite pulses, depression or potentiation.
Afterwards, cycles composed of 200 potentiation and 200 depression pulses are performed with changes in the pulse amplitudes and the pulse width in order to maximize the $On - Off$ ratio, the number of intermediate states and the stability upon repeated cycling.

Once the parameters allow repeated cycling with good stability, $On - Off$ ratio and number of intermediate states, a short symmetrization, composed of 1 potentiation and 1 depression pulse, is repeated 500 times. This symmetrization is performed to check if the device can find the symmetry point with the current parameters, a similar approach is shown in Figure 3.15. The symmetry point is ideally centered within the dynamic range of the device, but in practice it is usually skewed towards the HRS. A succesful symmetrization is stable, i.e. the value of the symmetry point does not change, and within $30 - 70\%$ of skew from the mean value of the dynamic range. If the symmetrization is not succesful, a further optimization of the pulse parameters is required.

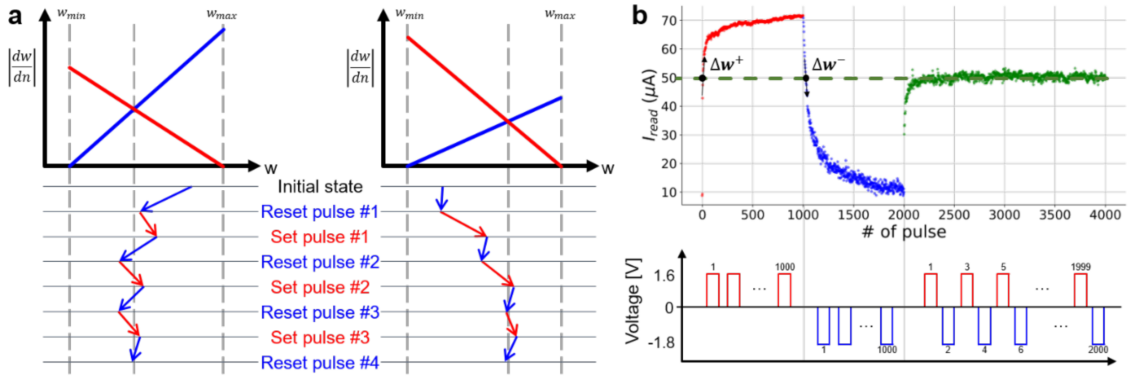

Figure 3.15: "[a] Procedure of finding the symmetry point with iterative set/reset pulses. Two exemplary cases of finding the symmetry point of each device are demonstrated. Note that both w_{sym} and the initial states are different in the two cases. As the device experiences set and reset pulses iteratively, the conductance states(i.e.weights) of the device tends to converge to the symmetry point. [b] Experimental data of potentiation, depression, and finding the symmetry point. When the device experiences set/reset pulses repeatedly, its state converges to the symmetry point where delta weights are equal for potentiation and depression. Corresponding voltage pulses are depicted below the graph."[21]

Once the optimal set of parameters is found, the performance evaluation can be started. The performance evaluation is based on test sequences composed of a defined number of cycles, performed by applying 200 potentiation and 200 depression pulses or vice versa, and the symmetrization, which is perfrormed by applying 1 potentiation and 1 depression pulse or vice versa, for a defined amount of pulses.
It is necessary to perform cycling before symmetrization because it is necessary to associate the symmetrization to a prior cycling to calculate some figures of merit of the symmetry point.
The number of cycles, the amount of pulses used for the symmetrization and the number of test sequences depend on the purpose of the test. A longer test with multiple repeats will give more information about long-term stability and endurance of the device, while a shorter test will only evaluate the current performance of the device.

3.2.3 Extraction of figures of merit

Cycling The HRS resistance of a cycling run is calculated by taking the median of the measured resistance for the last 10 depression pulses of each cycle and then averaging across all the cycles, Equation 3.3. The same procedure is applied for the LRS by considering the potentiation pulses, Equation 3.4.

$$R_{HRS\_AC} = \frac{1}{\#cycles} \cdot \sum_{i}^{\#cycles} median(R^i_{depression}[190 \rightarrow 200]) = \frac{1}{G_{HRS\_AC}} \tag{3.3}$$

$$R_{LRS\_AC} = \frac{1}{\#cycles} \cdot \sum_{i}^{\#cycles} median(R^i_{potentiation}[190 \rightarrow 200]) = \frac{1}{G_{LRS\_AC}} \tag{3.4}$$

The $On - Off$ ratio is defined as in Equation 3.5.

$$On - Off_{ratio} = \frac{R_{HRS\_AC}}{R_{LRS\_AC}} = \frac{G_{LRS\_AC}}{G_{HRS\_AC}} \tag{3.5}$$

The dynamic range, expressed in conductance units, is called G_{range} and it is defined as in Equation 3.6.

$$G_{range} = G_{LRS\_AC} - G_{HRS\_AC} \tag{3.6}$$

Symmetrization The symmetry point average resistance value is calculated as the mean of the resistance value during the whole symmetrization pulse sequence, Equation 3.7.

$$R_{SP} = \frac{2}{\#pulses} \cdot \sum_{i}^{\#pulses/2} \frac{R^i_{potentiation} + R^i_{depression}}{2} = \frac{1}{G_{SP}} \tag{3.7}$$

The change in conductance for each pulse during symmetrization ΔG^i_{SP} is defined in Equation 3.8.

$$\Delta G^i_{SP} = |G^i_{potentiation} - G^i_{depression}| \tag{3.8}$$

The average change in conductance ΔG^{mean}_{SP} can be calculated by averaging ΔG^i_{SP} over the whole symmetrization pulse sequence, Equation 3.9

$$\Delta G^{mean}_{SP} = \frac{2}{\#pulses} \cdot \sum_{i}^{\#pulses/2} \Delta G^i_{SP} \tag{3.9}$$

The symmetry point skew expresses the relative distance of the mean resistance/conductance of the symmetry point G_{SP} from the mean resistance/conductance of the HRS, Equation 3.10, and LRS, Equation 3.11.

$$SP_{skew\_HRS} = \frac{G_{SP} - G_{HRS\_AC}}{G_{range}} \tag{3.10}$$

$$SP_{skew\_LRS} = \frac{G_{LRS\_AC} - G_{SP}}{G_{range}} \tag{3.11}$$

The program noise indicates how consistent the change in resistance or conductance of the device is upon the application of a single pulse. This figure of merit is computed by calculating the standard deviation of ΔG^i_{SP} and normalizing it with respect to ΔG^{mean}_{SP}, as in Equation 3.12.

$$Program\ noise = \frac{1}{\Delta G^{mean}_{SP}} \cdot \sqrt{\frac{2}{\#pulses} \cdot \sum_{i}^{\#pulses/2} |\Delta G^i_{SP} - \Delta G^{mean}_{SP}|^2} \tag{3.12}$$

The number of states, which should quantify how many intermediate levels can be reached between HRS and LRS, is calculated by dividing the G_{range} by the ΔG^{mean}_{SP}, Equation 3.13. The number of states calculated with this definition is only relevant if the program noise is kept under control, eg. $< 150\%$. This definition of number of states is not perfect, since ΔG^i_{SP} can be lower with respect to $\max\{\Delta G^i_{cycling}\}$, but it is more conservative and robust against non-linear behaviour with respect to assuming 200 intermediate states resulting from 200 applied potentiation and depression pulses.

$$\#states = \frac{G_{range}}{\Delta G^{mean}_{SP}} \tag{3.13}$$

3.2.4 Fitting to a device model for AiHwKit

The IBM analog hardware acceleration kit [33] is an open source toolkit able to simulate the computations performed on crossbar arrays. The toolkit enables the creation of arbitrary ANNs using the PYTORCH framework and can simulate their implementation on crossbar arrays. The memristive elements in the crossbar arrays are emulated by numerous mathematical models which can consider also their non-idealities such as device-to-device variability, cycle-to-cycle variability, asymmetry and non linearity of the potentiation and depression curves, drift and noise.

Among the available mathematical models to describe the potentiation and depression curves the "PowStepDevice" is the closest to the tested devices. The value of the weight after a potentiation, weight increase, or depression, weight decrease, pulse in the "PowStepDevice" depends on the current weight and some parameters as detailed in Equations 3.14 and 3.15 respectively. Since a large scale AC testing would be required to evaluate precisely the variability of the devices the associated parameters have been initially neglected.

$$w_{ij}^{up} \leftarrow w_{ij} + dw_{up} \cdot \Omega_{ij}^{\gamma \cdot (1+\beta_{pow})} \tag{3.14}$$

$$w_{ij}^{dn} \leftarrow w_{ij} + dw_{dn} \cdot \Omega_{ij}^{-\gamma \cdot (1-\beta_{pow})} \tag{3.15}$$

$$\Omega_{ij} = \frac{(1 - w_{ij})}{2} \tag{3.16}$$

$$dw_{up} = dw_{min} \cdot (1 + \beta) \tag{3.17}$$

$$dw_{dn} = -dw_{min} \cdot (1 - \beta) \tag{3.18}$$

After normalizing the conductance range to the weight range, [-1,1], the parameters of the "PowStepDevice" model can be optimized to fit at best the experimental potentiation and depression curves. The fitting is performed by calculating the squared error of the "PowStepDevice" model with respect to the experimental potentiation and depression curves and minimizing it using the MATLAB function "fminsearch". Once the device is modeled crossbar arrays with the modeled memristor can be simulated.

3.3 Temperature dependent DC measurements

3.3.1 Setup

Temperature-dependent measurements aim at measuring the effect of elevated temperature on an RRAM device. For temperature variation of the set-up, the SUSS probe station is equipped with a cooling/heating unit, ATT systems P30, which can actively control the temperature inside the probe station and more precisely on the chuck, where the chip is placed. The cooling unit is able to force temperatures from 15 °C up to 200 °C. However, to avoid condensation on the chip the minimum temperature should be kept close to room temperature. The maximum temperature that the gold plated ceramic chuck can handle is 150 °C. Considering these temperature bounds, the temperature range used for the measurements is between 25 °C and 150 °C.

Due to the high temperature inside the probe station, the metal electrode pads on the devices can oxidize, thus influencing the outcome of the measurement. To combat this problem, a steady flux of nitrogen N_2 is blown inside the probe station chassis while the chip is heated, limiting the oxygen available and therefore slowing the oxidation process.

3.3.2 Experimental procedures

The maximum temperature that a memristive device can withstand reversibly with respect to its behavior at room-temperature is typically unknown for a novel material stack. Several processes can be influenced by temperature stress. For example, at elevated temperatures diffusion processes are favoured, which can critically influence the operation of the tested RRAM devices, since it relies on the diffusion of oxygen ions and vacancies. Additionally, heating and cooling a sample can also build stress between the various layers of the material stack due to their different thermal expansion coefficient, thus altering the material properties such as the mobility [24].

For these reasons, temperature-dependent experiments should be carefully planned. The first parameters to choose are the temperature steps and the temperature range. Afterwards, a measurement plan should be made to determine which kind of measurement will be performed on which device and at which temperature. An important factor that has to be taken into account is the permanence of the chip at the tested temperature for the overall duration of the experiment. This is important for two reasons: the history of temperatures seen by the chip and the fact that the measurements should be performed continuously. Since it is assumed that temperature variation is a non-reversible process, the approach followed for performing temperature-dependent measurements is detailed as follows:

- Room temperature characterization of all parts of the chip not involved in temperature-dependent experiments (Time: anytime before T-dep experiment)
- Room temperature preparation of devices for tempearture-dependent experiment (Time: as close as possible to the T-dep experiment)
- Temperature-dependent experiment (chip must not be moved from the chuck and the measurements should be performed continuously)
  - If some device have been prepared for the experiment, check the condition of those devices
  - Set Temperature 1 and wait 10 min to reach temperature steady state in the probe station
  - Temperature 1: perform all measurements planned for Temperature 1
  - Set Temperature 2 (Temperature 2 $>$ Temperature 1) and wait 10 min to reach temperature steady state in the probe station
  - Temperature 2: perform all measurements planned for Temperature 2
  - Repeat until last planned temperature
  - Cool down gradually to T=25 °C
  - Measure again at T=25 °C to see if there are any permanent changes in the device behavior

With this procedure, the device characterization is always performed for monotonically increasing temperatures.

There are multiple experiments that can be performed during a temperature-dependent measurement run. In particular, the temperature dependence of the following states and processes can be studied:

- the leakage current
- the forming process
- the HRS and LRS resistance of a device previously formed and cycled
- the cycling characteristics of a device previously formed and cycled
- the cycling characteristics of a device formed at the same temperature

- the cycling characteristics at room temperature for a device formed at elevated temperature

3.3.3 Conduction Mechanisms analysis

The testing methodologies briefly summarized in the previous sections are of interest also to understand the physical mechanisms at microscopic level for various device types, including the studied memristive devices. The electrical I-V characteristics, as well as the response of the I-V behavior to elevated temperature stress, can be used to extract physical parameters (i.e., layer thickness, permittivity, etc..), which can furthermore be double-checked by other inspection methodologies (i.e., microscopic inspection like SEM and TEM, ellipsometer, etc. . .) to clarify the energy band diagram and to model the conduction mechanism(s) within the various layers of the device.
Extensive literature can be found on the physical mechanisms of charge transport in dielectrics. This is due to the use of dielectric materials in common electrical devices (e.g., MOSFET, DRAMs, Flash memories) and the urgence to precisely characterize such dielectric layers for device behaviour, modeling, and reliability reasons. The conduction mechanisms responsible for the leakage current in a given structure depend on the particular material stack and the device geometry. In the studied memristive devices, the leakage current is defined as the current flowing through the device before any irreversible resistance state change (i.e., soft or hard breakdown of the dielectric layer) has ever happened. This definition results in studying the IV characteristic for $V < V_{forming}$. In the following, the experimental response as well as some conduction mechanism considerations for metal-insulator-metal (MIM) structures and metal-insulator-semiconductor (MIS) structures are discussed. The MIM structure enclose the dielectric in between two metal contacts. The other kind of structure, the MIS, is intrinsically asymmetric since one electrode is a metal while the other is a semiconductor. This structure resembles more closely the actual electronic devices, but the voltage drop across the semiconductor should be evaluated carefully and subtracted from the overall voltage applied in order to obtain the voltage drop across the insulator [7].

The conduction mechanisms can be grouped into two classes: the electrode-limited conduction mechanisms and the bulk-limited conduction mechanisms, Figure 3.16,[22].

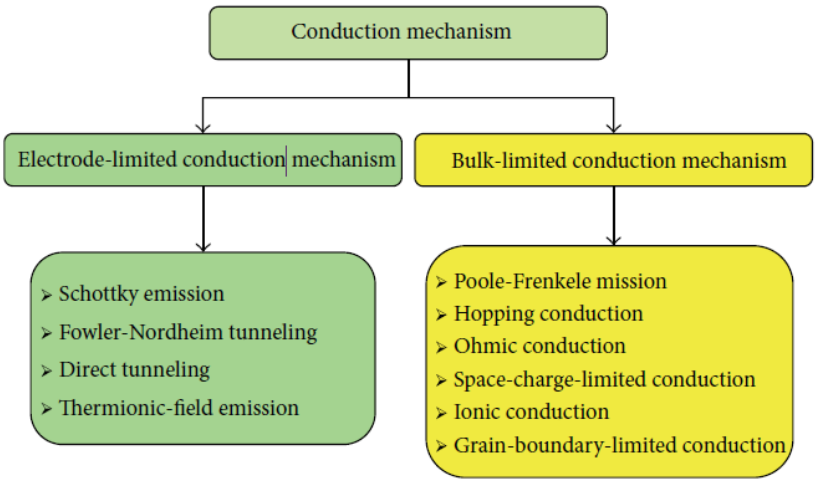

Figure 3.16: "Classification of conduction mechanisms in dielectric films."[7]

Electrode-limited conduction mechanisms mainly depend on the properties of the electrode-dielectric interface, such as the barrier height. The electrode-limited conduction mechanisms comprehend Fowler-Nordheim tunneling, Schottky or thermionic emission, thermionic-field emission, and direct tunneling [7]. In Figure 3.17, a scheme of the electrical charge transport across a sequence of layers is shown.
Bulk-limited conduction methods are influenced by the properties of the insulator material, for this

reason the most important parameters are the density and the energy position of the defects in the dielectric. The bulk-limited conduction mechanisms include: Ohmic conduction, hopping conduction, Poole-Frenkel emission, space-charge-limited conduction, grain-boundary-limited conduction, ionic conduction and multiphonon trap-assisted tunneling[7].

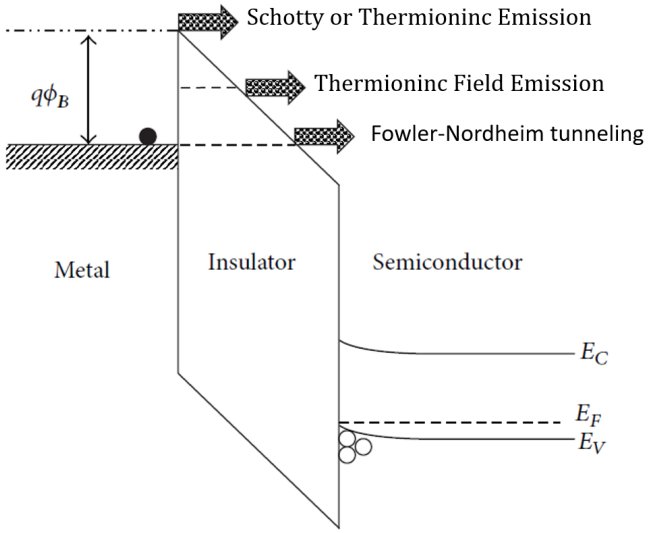

Figure 3.17: Comparison of Schottky or thermionic emission, thermioninc field emission and Fowler-Nordheim tunneling, adapted from[7].

Schotty or Thermioninc Emission is a conduction mechanism where the electrons overcome the energy barrier from the metal work function to the bottom of the dielectric conduction band by means of thermal energy. Therefore this mechanism is mostly observed at high temperatures. The current density J due to the thermionic emission can be expressed as in Equation 3.19.

$$J = A^* T^2 \exp\left[\frac{-q\left(\phi_B - \sqrt{\frac{qE}{4\pi\epsilon_r\epsilon_0}}\right)}{k_B T}\right], \tag{3.19}$$

$$A^* = \frac{4\pi q k_B^2 m^*}{h^3} = \frac{120 m^*}{m_0}$$

Where "A^* is the effective Richardson constant, m_0 is the free electron mass, m^* is the effective electron mass in the dielectric, T is the absolute temperature, q is the electronic charge, $q\phi_B$ is the Schottky barrier height (i.e. conduction band offset), E is the electric field across the dielectric, k_B is the Boltzmann's constant, h is the Planck's constant, ϵ_0 is the permittivity in vacuum, and ϵ_r is the optical dielectric constant (i.e. the dynamic dielectric constant)" [7].

Fowler-Nordheim tunneling When the applied electric field is high enough, the potential barrier seen by the electrons in the electrode is not rectangular anymore, but it is triangular. Thus, the width of the barrier is reduced favouring tunneling to the conduction band of the dielectric. Equation 3.20 expresses the current density due to the Foweler-Nordheim tunneling [7].

$$J = \frac{q^3 E^2}{8\pi h q \phi_B} exp\left[\frac{-8\pi\sqrt{2qm_T^*}}{3hE}\phi_B^{3/2}\right] \tag{3.20}$$

"Where m_T^* is the tunneling effective mass in the dielectric" [7].

Direct Tunneling The direct tunneling mechanism is relevant only when the thickness(es) of the dielectric layer(s) are very small, eg. $< 3.5\,\text{nm}$ [25], therefore it will not be taken into account in this analysis.

Thermionic-Field Emission is relevant at intermediate biases from thermionic emission and FN tunneling. It considers electrons with energies between the Fermi level of the metal and the bottom of the conduction band of the dielectric. The current density due to this mechanism can be expressed as in Equation 3.21 [15].

$$J = \frac{q^2\sqrt{m_0 k_B T}E}{8\hbar\pi^{5/2}} exp\left(\frac{q\phi_B}{k_B T}\right) \exp\left[\frac{\hbar^2 q^2 E^2}{24 m_0 \left(k_B T\right)^3}\right] \tag{3.21}$$

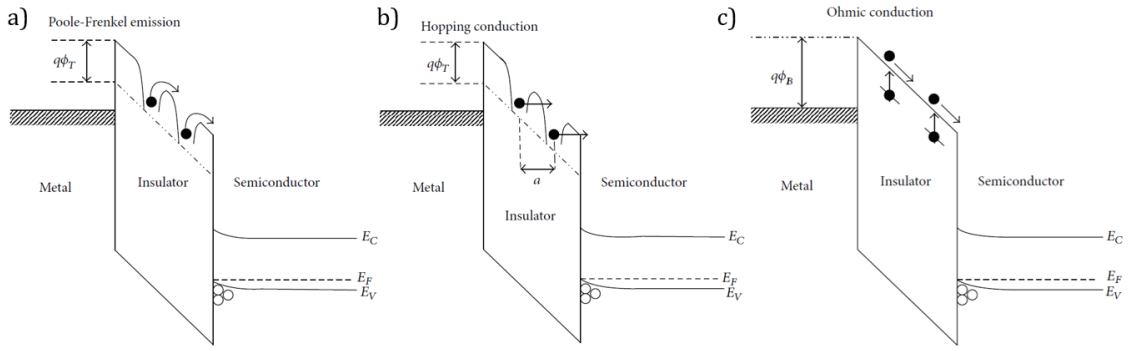

Figure 3.18: Comparison of Poole-Frenkel emission a), Hopping conduction b), and Ohmic conduction c), adapted from[7].

Poole-Frenkel Emission is a mechanism in which the carriers are thermally excited to the conduction band of the dielectric from the trap sites. This transition can be energetically favourable since it can reduce the Coulomb potential energy due to repulsion of neighbouring electrons. This mechanism is mostly observed at high temperature and high electric field, since the carriers need to be excited to the conduction band of the dielectric [9]. The current density due to Poole-Frenkel emission can be expressed as in Equation 3.22 [7].

$$J = q\mu N_C E\, exp\left[\frac{-q\left(\phi_T - \sqrt{\frac{qE}{\pi\epsilon_r\epsilon_0}}\right)}{k_B T}\right] \tag{3.22}$$

"Where μ is the electronic drift mobility, N_C is the density of states in the conduction band, and $q\phi_T$ is the" distance of the trap energy level from the bottom of the conduction band of the dielectric [7].

Hopping Conduction In the Hopping conduction mechanism the conduction is due to tunneling of trapped electrons between the trap sites in the dielectric. The current density due to this mechanism is written in Equation 3.23 [29].

$$J = qan\nu\, exp\left[\frac{qaE}{k_B T} - \frac{q\phi_T}{k_B T}\right] \tag{3.23}$$

"Where a is the mean hopping distance (i.e. the mean spacing between trap sites), n is the electron concentration in the conduction band of the dielectric, and ν is the frequency of thermal vibration

of electrons at trap sites", and $q\phi_T$ is the distance of the trap energy level from the bottom of the conduction band of the dielectric [7].

Ohmic Conduction is due to the drift of carriers in the conduction or valence band of the dielectric. Even though the dielectric layer has a high band gap, a small concentration of free carriers in the band is present due to thermal excitation of electrons from the tails of the Maxwell-Boltzmann distribution. Since the amount of free carriers is very small, Ohmic conduction can be observed only at very low biases when there are no other significant contributions. The current density due to Ohmic conduction can be expressed as in Equation 3.24 [7].

$$J = \sigma E = nq\mu E\,, \tag{3.24}$$

$$n = N_C exp\left[\frac{-(E_C - E_F)}{k_B T}\right]$$

Where σ is the electrical conductivity.

Space-charge-limited conduction is a model which describes the conduction splitting in in three different regions, as shown in Figure 3.19: the Ohmic region ($J_{Ohmic} \alpha V$), the trap-filled limited region ($J_{TFL} \alpha V^2$), and the Child's law region ($J_{Child} \alpha V^2$). The transition voltage from the Ohmic to the trap-filled limited region is V_{tr}, while the transition voltage from the trap-filled limited region to the Child's law region is V_{TFL}. Equations 3.25, 3.26 and 3.27 describe the current density in the three regions, while equations 3.28 and 3.29 describe the transition voltages [7].

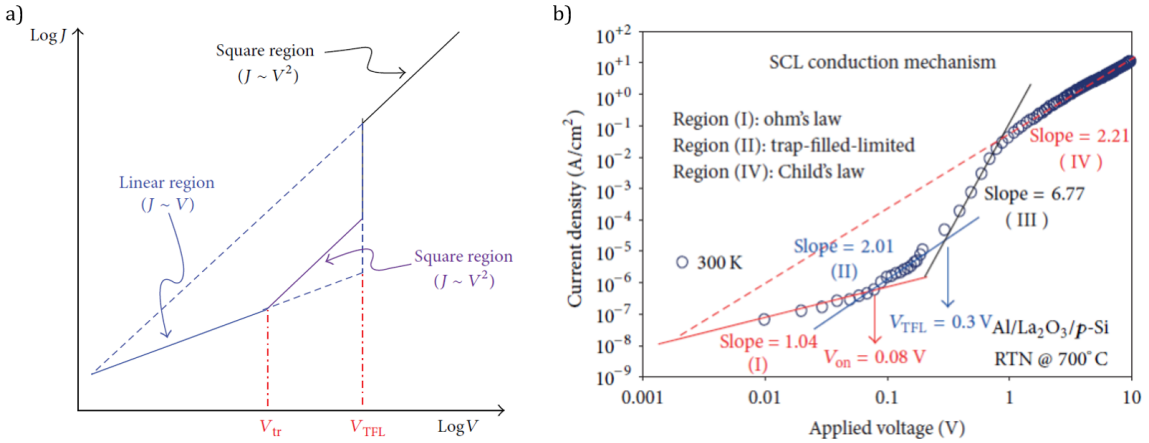

Figure 3.19: Typical current density-voltage (J-V) characteristic of space-charge-limited conduction current in the ideal case a), and an experimental curve b); V_{tr} is transition voltage, V_{TFL} is trap-filled limit voltage; adapted from [7].

$$J_{Ohmic} = q\nu_0\mu\frac{V}{d} \tag{3.25}$$

$$J_{TFL} = \frac{9}{8}\mu\epsilon\theta\frac{V^2}{d^3} \tag{3.26}$$

$$J_{Child} = \frac{9}{8}\mu\epsilon\frac{V^2}{d^3} \tag{3.27}$$

$$V_{tr} = \frac{8}{9}\frac{q\nu_0 d^2}{\epsilon\theta} \tag{3.28}$$

$$V_{TFL} = \frac{qN_t d^2}{2\epsilon} \tag{3.29}$$

Where d is the thickness of the dielectric, ϵ is the static dielectric constant, N_t is the trap density and θ is [7]:

$$\theta = \frac{N_C}{g_n N_t} exp\left[\frac{E_t - E_C}{k_B T}\right],$$

where E_t is the trap energy level and g_n is the degeneracy of the energy states in the conduction band of the dielectric.

When the applied bias is small, the condition of weak injection is reached: the amount of thermally generated free carriers (ν_0) is much bigger than the injected carriers [8], thus Ohmic conduction is observed.
When $V > V_{tr}$ the dielectric is in strong injection an the contribution to the current due to the injected carriers flowing into the traps has to be taken into account. In this condition, the traps are partially filled and the conduction is trap-fill limited.
"While the bias voltage reaches V_{TFL} in the strong injection mode, the traps get gradually saturated, which means that the Fermi-level gets closer to the bottom of the conduction band. This results in a strong increase of the number of free electrons, thus explaining the increase of the current for $V = V_{TFL}$" [7].
After V_{TFL}, there is very strong injection and all the traps are filled. "Thus a space charge layer in the dielectric builds up" and "the electric field cannot be regarded as constant any longer and the current is fully controlled by the space charge, which limits the further injection of free carriers in the dielectric. Square law dependence of the current (Child's law) is the consequence of the space charge controlled current" [7].

Multiphonon trap-assisted tunneling is a model in which the conduction is described by tunneling percolation paths of carriers in the dielectric's traps. The capture and emission from the electrodes, as well as the tunneling between the traps can be assisted by the excitation or absorption of multiple phonons. Since this model relies on percolation paths made by defects, a model representing the distribution in space and energy of the defects in the dielectric should be made. All possible percolation paths between the two electrodes should be considered and summed up.
For each percolation path the associated current is limited by the slowest trap to capture and emit electrons in the path, thus: [23]

$$I_{path} = \frac{q}{\tau_{c,max} + \tau_{e,max}} \tag{3.30}$$

where $\tau_{c,max}$ and $\tau_{e,max}$ are the time constants related to the capture and emission of electrons in the slowest defect of the path.
Considering phonons with a single frequency ω_0 and indicating with m and n the number of phonons involved in the emission and capture process respectively, the time constants related to the trap j in the percolation path $\tau_{c,j}$ and $\tau_{e,j}$ can be expressed as in Equations 3.31 and 3.32 [35].

$$\tau_{c,j}^{-1} = \sum_m (\tau_{c,j,m})^{-1} \tag{3.31}$$

$$\tau_{e,j}^{-1} = \sum_n (\tau_{e,j,n})^{-1} \tag{3.32}$$

$\tau_{c,j,m}$ is "the time required for the e^-/h^+ capture into the trap j of the conduction path, associated with the release of the e^-/h^+ energy $m\hbar\omega_0$ to the lattice, with m being the number of phonons involved", Equation 3.33 [35].
$\tau_{e,j,m}$ is "the time required for the e^-/h^+ emission from the trap j of the conduction path, associated with the absorption of the e^-/h^+ energy $n\hbar\omega_0$ from the lattice, with n being the number of phonons involved", Equation 3.34 [35]. Figure 3.20 depicts an example of multiphonon assisted e^-/h^+ capture from a semiconductor to a trap and the subsequent multiphonon assisted e^-/h^+ emission from a trap to the conduction band of a metal.

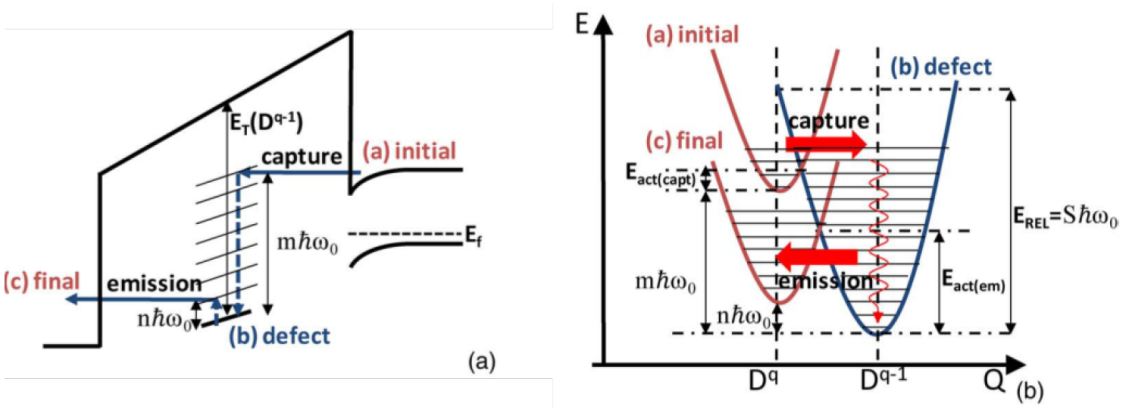

Figure 3.20: "a) Band diagram illustrating an example of multiphonon capture and emission processes in the case of a single-trap conductive path. A monolayer stack has been considered for simplicity. b) Full energy configuration coordination diagram corresponding to the processes" [35].

$$\tau_{c,j,m}^{-1} = N_{j-1}\left(E_j + m\hbar\omega_0\right) f_{j-1}\left(E_j + m\hbar\omega_0\right) C_{j,m} P_T\left(E_{j-1}, E_j + m\hbar\omega_0\right) \tag{3.33}$$

$$\tau_{e,j,m}^{-1} = N_{j+1}\left(E_j + n\hbar\omega_0\right) \cdot \left(1 - f_{j+1}\left(E_j + n\hbar\omega_0\right)\right) Em_{j,n} P_T\left(E_j + n\hbar\omega_0, E_{j+1}\right) \tag{3.34}$$

Where E_j is the energy level of the trap j; N_j and f_j are the density of states and the Fermi-Dirac occupation probability of either the previous trap or the cathode, for Equation 3.33, or the next trap or the anode, for Equation 3.34; P_T is the tunneling probability which can be calculated using the Wenzel-Kramer-Brillouin (WKB) approximation as in Equation 3.35 [39], where r_j is the position of the defect j;

$$P_T = exp\left[-\frac{2}{\hbar}\int_{r_j}^{r_{j+1}} \sqrt{2m_T^*\left(E_{j+1} - E_j\right)}dr\right] \tag{3.35}$$

C_j, m and Em_j, n are "the trap capture and emission rates that account for the carrier–phonon interaction" as expressed in [35].

3.4 Ginestra

Ginestra is a software environment developed by Applied Materials. It allows simulation of the physical and electrical operation of electronic devices such as FinFET, FeFET, RRAM, FeRAM and PCM. In particular, the software has an advanced physics model able to precisely describe the charge/ion transport, the charge trapping and detrapping, the material modifications due to phenomena such as polarization switching, phase change, degradation, and more [1]. On top of that there is the standard electromagnetic, mechanical and thermal simulation.

3.4.1 Device model

The first step to use Ginestra is to build a device model from the materials present in the Applied Material database or by creating a custom material. The dimensions, the material stack and the kind of electrodes have to be specified. In addition, it is possible to add species in the layers of the device which can model e.g. oxygen ions and vacancies. The species can be placed in a material with a defined space and energy distribution, they can be generated, they can recombine and they can also drift across different materials if it is allowed for the simulation.

3.4.2 Simulations

In the Ginestra software numerous kinds of simulations can be performed, however only the kinds used in this work will be covered.

Design of experiment This kind of simulation enables current voltage I-V and capacitance voltage C-V simulations on the selected device. The simulations can be dynamic or static.
Dynamic simulations take the effect of a voltage sweep on the device into account. This allows following the device response to the voltage sweep in time, which is useful for simulating the forming and the set/reset operations, where oxygen vacancies and ions are generated and moved.
Static simulations do not consider any device change throughout the voltage sweep, therefore the device response is evaluated independently for each voltage step. These simulations are much faster than dynamic simulations and are used for the pristine sweeps since it is assumed that the device does not change.
For both kinds of simulations it is possible to change one or multiple material parameters to a set of custom values in order to evaluate their impact on the simulation output.

Parametric extraction The parametric extraction tool allows changes one or multiple specified material parameters, within certain bounds, to make the simulation output as close as possible to the provided I-V or C-V experimental data. This tool works by performing static simulation at each voltage point, therefore it can be used only when the device characteristics do not change, i.e. during pristine sweeps. To optimize the parameters, the software uses multiple optimization algorithms which can be tuned independently. In particular it is possible to specify if it is required to search for a global minimum or just a local minimum and how many maximum iterations are possible for both the global optimization and the local optimization routine.

Physical options For all the simulations it is possible to set the temperatures at which they are to be performed and to specify which parts of the physical solver should be active. It is possible to enable or disable the electron transport in the Conduction band, the hole transport in the Valence band, the transport in sub-bands created by defects, direct tunneling (electrode to electrode), tunneling between electrode and Conduction band, tunneling between electrode and Valence band, intra-band tunneling for both the Conduction and Valence bands, Valence band to Conduction band tunneling, trap to electrode electron or hole transitions, trap to trap electron or hole transitions, Conduction and Valence band electron or hole transitions, the recursive solution of the Poisson's equation in order to account for doping, fixed charges and defects/ion charges, the solution of the Fourier's equation for heat transfer in order to have the exact calculation of the local temperature in the device due to Joule heating and dissipation and the generation and recombination of defects. For simulations of resistive RAMs, all the transport mechanisms are activated. The generation, recombination and the Fourier's equation solution are only enabled for forming and Set/Reset transitions, since in pristine sweeps is assumed that there is no generation of defects.

4 Results and analysis

4.1 Comparison between $HfO_2 - Ti$ and $HfO_2 - MO_x$ devices

The aim of this section is to compare the electrical behavior of the well-known $HfO_2 - Ti$ technology, refered to as baseline, with the novel material stack, $HfO_2 - MO_x$, developed at IBM Research Zurich.

4.1.1 Forming

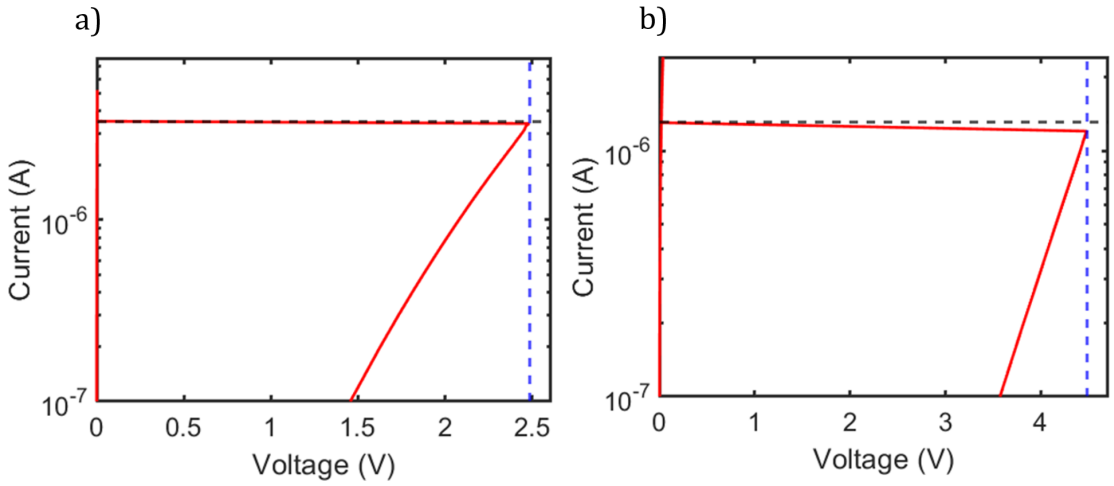

Figure 4.1: Representative current forming I-V curves of a) $HfO_2 - Ti$ technology, and b) $HfO_2 - MO_x$ technology. The blue dashed line indicates the forming voltage V_{form}, while the black dashed line indicates the forming current I_{form}. Device size: 12 µm; ramp steps: 100 nA; ramp rate: 10 steps/s.

In Figure 4.1, representative current forming I-V curves are depicted for the two analyzed RRAM technologies. For both cases a current ramp is applied with 100 nA steps on a 12 µm device.
As soon as the applied current reaches a critical value, the voltage suddenly "snaps back", indicating a dramatic decrease in the resistance of the device.
Already from this example it is clear that the baseline requires a much smaller voltage to complete the forming process. Moreover, if the voltage across the device prior to the forming event is compared at a given current for the technologies, it can be stated that the $HfO_2 - Ti$ device has a lower resistance with respect to the $HfO_2 - MO_x$ device. The larger pristine resistance in $HfO_2 - MO_x$ devices is attributed to the non-metallic characteristic of the MO_x layer, contrary to the Ti layer in $HfO_2 - Ti$ devices.

By analyzing the figures of merit extracted from multiple forming processes, it is evident that the forming voltage, V_{form}, increases with decreasing device area, Figure 4.2. This trend, observed for both technologies, can be explained by the fact that in a bigger device it is statistically more probable to find a larger cluster of defects. This cluster of defects reduces the effective thickness of the HfO_2 layer, thus locally increasing the electric field. A higher electric field induces more current, which in turn leads to local heating and therefore a higher generation rate of new defects [30]. The local generation of new defects starts a fast positive feedback which leads to the breakdown of the dielectric by formation of a conductive filament. Therefore, in large devices the forming event can be generated at lower applied voltages.

The forming current, I_{form}, increases with increasing device area, Figure 4.3, since in larger devices

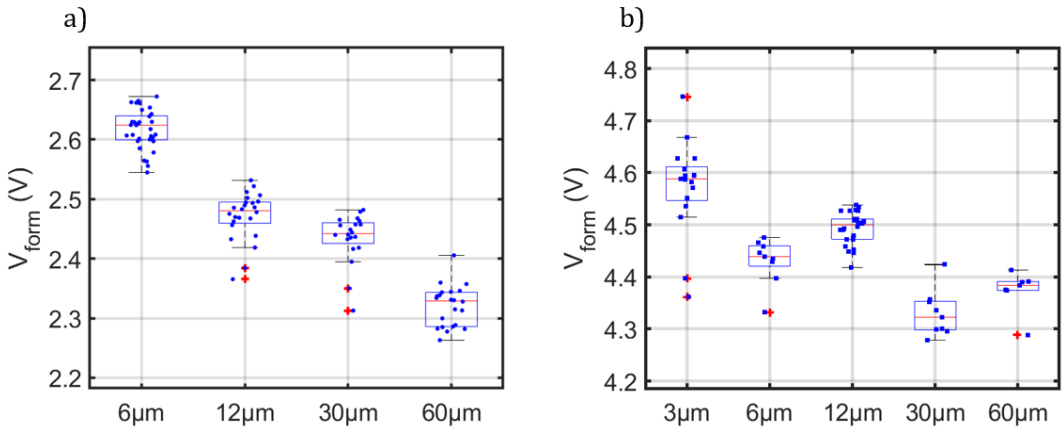

Figure 4.2: Boxplots representing the forming voltage, V_{form}, on multiple devices grouped by device size for a) the $HfO_2 - Ti$ technology, and b) the $HfO_2 - MO_x$ technology. The box encloses data between the 25th and 75th percentiles; the whiskers extend from the box edges up to 1.5 times the interquartile range. Red crosses represent data outside the whiskers. The blue dots are the actual data points, with a slight horizontal spread to better show the number of data points. The same conventions are kept for all subsequent boxplots.

there are more conductive paths through the dielectric. Thus, to force the same voltage, a larger current is required to trigger the forming event.

The current density, Figure 4.4, is obtained by normalizing the forming current to the device area, J_{form}. As expected, the current density is almost constant throughout the range of device sizes, with the exception of a small decrease with device size increase. Two hypotheses can be made to explain this observation: the first is that since in larger devices the forming voltage is lower, a smaller current density is required to force the forming voltage. The second hypothesis relies on the fact that the conduction is not uniform throughout the device area, but it mainly follows a preferential path, especially immediately before the forming event. The part of current flowing through this preferential path is not related to the area of the device. Therefore, the current density is smaller in larger devices, since a part of the current does not scale with the device area.

From Figures 4.1 and 4.2 it is evident that the $HfO_2 - MO_x$ technology requires an applied voltage of approximately 2 V higher than the $HfO_2 - Ti$ technology to form. This observation can likely be explained by two, probably concurring, effects. The first effect relates to the fact that the partially oxidized transition metal oxide cannot be assumed to be an ideal conductor, like Ti. Thus, there is a voltage drop on the MO_x layer. This effect cannot be the sole explanation, since the MO_x layer has a measured resistivity in the order of 1 Ω cm, which leads to a small voltage drop, below 100 mV, over this layer.
The second effect is related to the differences in the interfaces between HfO_2 and Ti or HfO_2 and MO_x. In the case of Ti, it is well known [3] that oxygen from the HfO_2 oxidizes the Ti at the interface to form a TiO_x thin film. The oxidation of Ti leads to a higher concentration of defects, oxygen vacancies, in the HfO_2 at the interface.
Since MO_x is a partially oxidized metal, it does not attract oxygen ions from HfO_2. Instead, it can be reduced by HfO_2. The formation of HfO_2 is more favorable than the formation of transition metal oxides, as reported in Figure 4.5. Therefore, at the interface there is a reduction of MO_x and an oxidation of HfO_2. Oxidation of the HfO_2 layer means that it becomes more stoichiometric and the number of defects is reduced. A lower amount of defects in the HfO_2 layer in the $HfO_2 - MO_x$ samples can explain the larger forming voltages and the smaller currents.

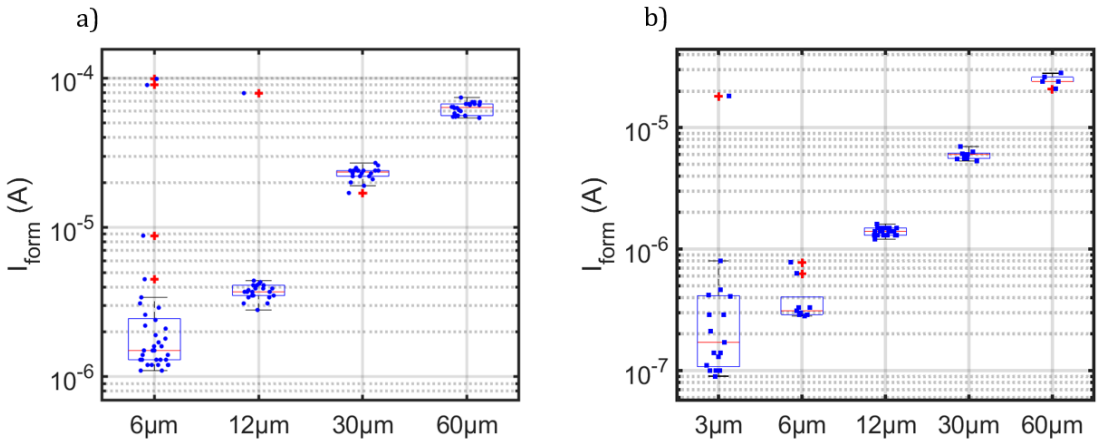

Figure 4.3: Boxplots representing the forming current, I_{form}, on multiple devices grouped by device size for a) the $HfO_2 - Ti$ technology, and b) the $HfO_2 - MO_x$ technology.

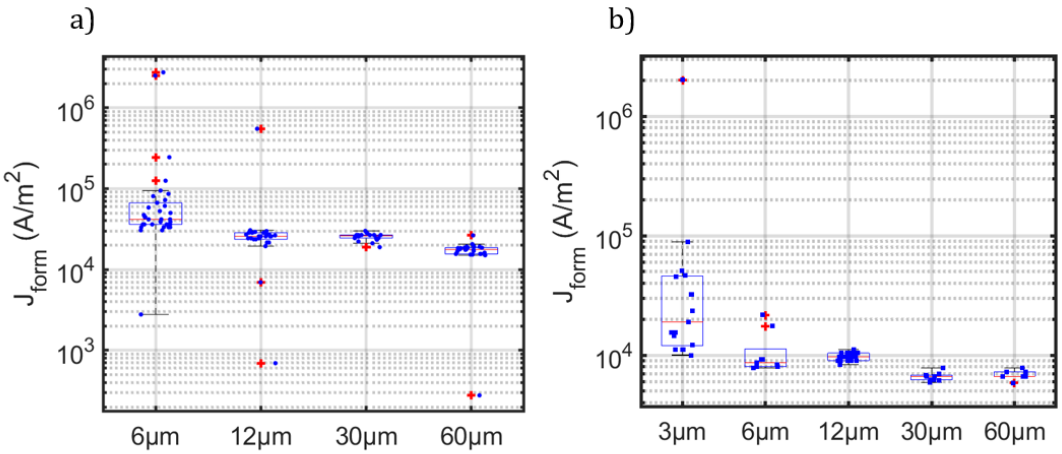

Figure 4.4: Boxplots representing the forming current density, J_{form}, on multiple devices grouped by device size for a) the $HfO_2 - Ti$ technology, and b) the $HfO_2 - MO_x$ technology.

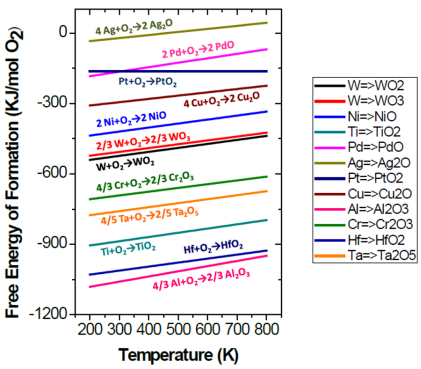

Figure 4.5: Ellingham diagram for various transition metals sub-oxides reactions and other metal oxides [6].

4.1.2 Post-forming resistance

After the forming event a conductive filament of oxygen vacancies is generated in the HfO_2 layer. The oxygen ions extracted during the forming process drift towards the top electrode, where a positive bias is applied. Thus, the oxygen ions accumulate in the above Ti or MO_x. In the TiO_x/Ti layer, an oxygen reservoir is formed, while instead in the MO_x layer an oxidized region on top of the filament is formed [5], this difference will lead to a different switching behavior for the technologies. Assuming a comparable filament resistance, the above statements can be confirmed by the post-forming resistance, defines as the resistance of the device after the forming event, monitored at 0.2 V, Figure 4.6. Because the post-forming resistance is much smaller for $HfO_2 - Ti$ devices with respect to $HfO_2 - MO_x$ devices, there is indeed a larger resistance associated with MO_x than with TiO_x/Ti.

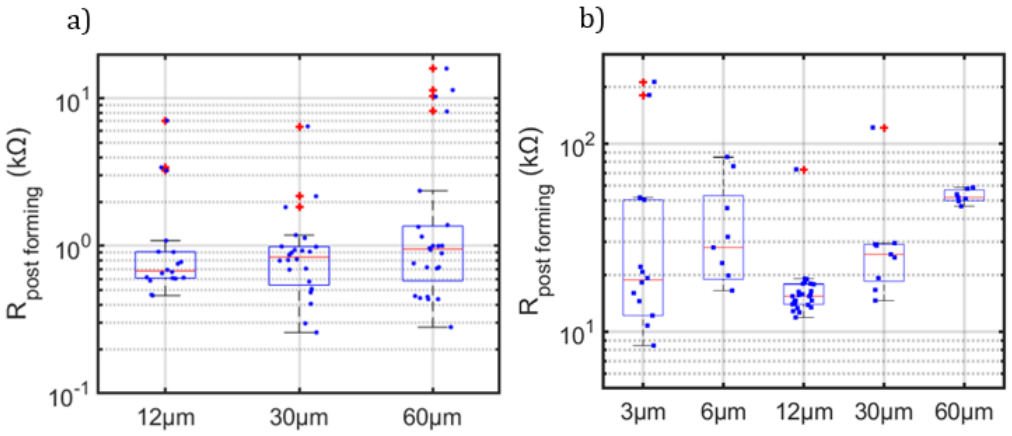

Figure 4.6: Boxplots representing the post-forming resistance, $R_{postforming}$, on multiple devices grouped by device size for a) the $HfO_2 - Ti$ technology, and b) the $HfO_2 - MO_x$ technology.

4.1.3 Switching mechanism

The two technologies need different procedures before cycling with fixed V_{stops}, as discussed in Section 3.1.4, because the physics behind the switching mechanism is different.

In the baseline technology the set transition is performed by applying a positive bias to the top electrode. The positive voltage attracts the generated oxygen ions in the scavenging layer, TiO_x/Ti, thus favouring the enlargement, or the formation, of the conductive filament.
The reset transition instead happens when a negative voltage is forced on the top electrode. This situation can also be viewed in the opposite way, thus a positive voltage is forced on the bottom electrode, while the top electrode is at the reference potential. A positive bias from the bottom to the top electrode causes the oxygen ions to drift from the TiO_x/Ti reservoir to the HfO_2 layer. There, the oxygen ions recombine with the oxygen vacancies, thus reducing the size of the conductive filament until it is interrupted. A more detailed explanation is has been reported in Section 2.

For the $HfO_2 - MO_x$ RRAM technology, the proposed switching mechanism is centered around a highly oxidized region in the MO_x which forms on top of the filament after forming. Supposedly, the resistance of this region is higher than the filament resistance and therefore, this region is responsible for the resistive switching.
After forming, Figure 2.3b, this region thus has the highest resistance. Then, when a Set operation with a high voltage is performed, the oxidized region decreases in volume and therefore the oxygen ions from the oxidized region are pushed towards the HfO_2 layer. For this operation, a negative bias

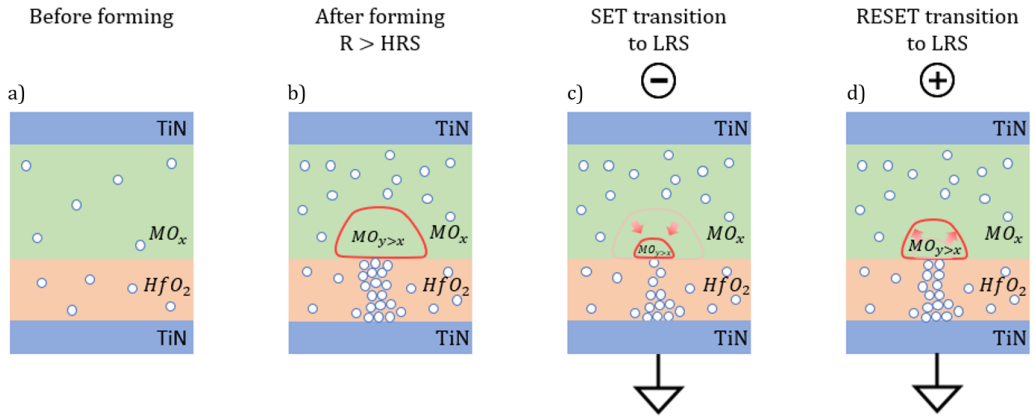

Figure 4.7: Schematic representation of the proposed switching mechanism for $HfO_2 - MO_x$ devices. The white dots represent oxygen vacancies, while the plus and the minus, in d) and c) respectively, indicate that a positive or negative bias is applied from the top to the bottom electrode.

between the top and bottom electrode is required, Figure 2.3c. Even though the size of the filament in the HfO_2 layer decreases, i.e. its resistance increases, the overall resistance of the device decreases because the resistance of the MO_x layer decreases and the filament remains intact.
The Reset operation is performed with a positive voltage bias. Oxygen ions are generated in the HfO_2 layer and drift to the MO_x layer, which enlarges the oxidized region. Hence, the resistance of the RRAM device increases. During this operation the conductive filament in HfO_2 layer becomes thicker. However, the oxidized region in the MO_x layer provides the dominant contribution to the resistance of the RRAM device, thus the overall resisrance increases. The Reset operation is performed with a smaller voltage bias compared to forming, because the device is intially in a lower resistance state than the pristine state. Therefore, more current can flow, causing a temperature increase which favours the generation of oxygen ion-vacancy pairs. At the same time the number of oxygen ions in the MO_x layer is smaller compared to the forming event. Therefore also the more oxidized region is smaller. Consequently, the post-forming resistance is the upper bound of the HRS resistance as confirmed by the data in Figure 4.8.

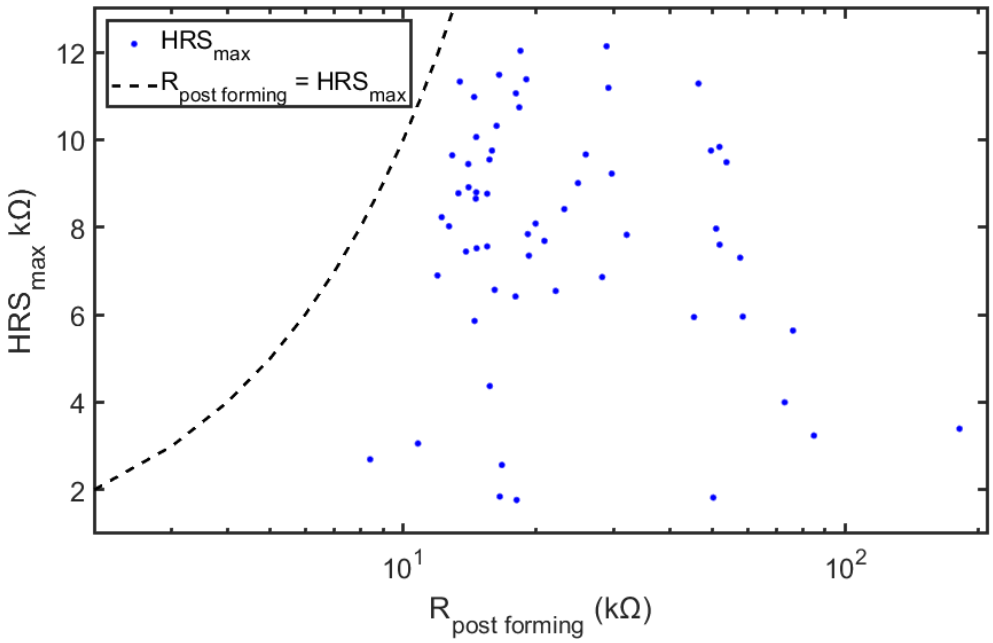

Figure 4.8: Maximum HRS value reached by the device, HRS_{max}, as a function of the post-forming resistance, $R_{post-forming}$. Blue dots are experimental data points and the black line indicates the values for which $R_{post-forming} = HRS_{max}$.

4.1.4 DC cycling

After the "wake up"/stabilization procedure, Section 3.1.4, the devices can be cycled between fixed V_{stops} and it is possible to evaluate their performances in terms of HRS and LRS values, On-Off ratio, abruptness of the transitions, cycle-to-cycle variability and stability upon repeated cycling.
In Figure 4.9, a comparison between single baseline and $HfO_2 - MO_x$ devices is shown. The devices are cycled ten times and the results are shown as I-V plots, R-V plots and plots showing the HRS and LRS resistance values across the cycles. From the I-V and R-V curves, Figures 4.9 [a,b,d,e], the difference in the polarity between the devices, as explained in the previous subsection, can be observed: the $HfO_2 - Ti$ devices set at positive voltages and reset at negative voltages, while the $HfO_2 - MO_x$ devices show the opposite behaviour.
It can also be seen that the $HfO_2 - MO_x$ technology shows smoother Set/Reset transitions and lower cycle-to-cycle variability, but it has a smaller HRS/LRS ratio.

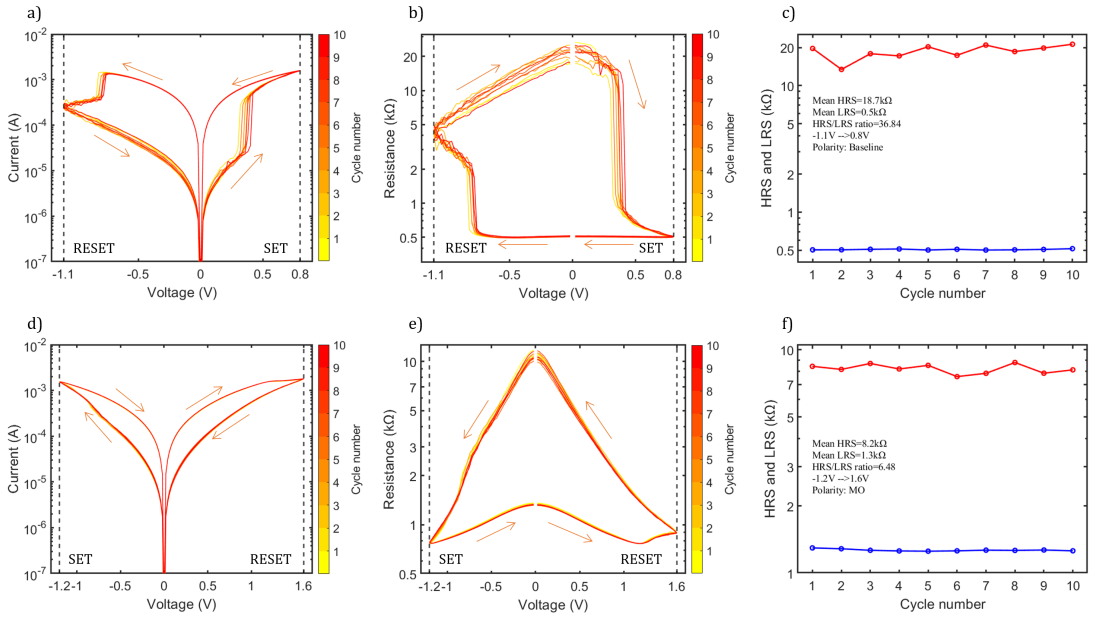

Figure 4.9: IV, current-voltage, [a,d] curve, RV, resistance-voltage, [b,e] curve, and recorded HRS-LRS values [c,f], for [a,b,c] the $HfO_2 - Ti$ technology, and [d,e,f] the $HfO_2 - MO_x$ technology.

HRS and LRS resistance The HRS and LRS resistances are comparable for the two technologies, however it is evident from Figure 4.10 that the baseline technology has an LRS with a consistently lower resistance and a tighter distribution across multiple devices.
In LRS, the baseline devices have a large conductive filament with on top the conductive Ti layer, whereas the $HfO_2 - MO_x$ devices have a smaller filament with on top a small region of more oxidized MO_x plus the remaining MO_x, Figure 4.7. Thus, the MO_x layer is less conductive than the Ti layer. Therefore, the lower LRS resistance in the $HfO_2 - Ti$ with respect to the $HfO_2 - MO_x$ technology is consistent with the proposed model for the $HfO_2 - MO_x$ devices. Moreover, the higher variability in the LRS of the $HfO_2 - MO_x$ device can be mostly attributed to the variability in the more oxidized region of MO_x on top of the filament. The extremely low variability, both device-to-device, Figure 4.10 a, and cycle-to-cycle, Figure 4.13 b, in the $HfO_2 - Ti$ devices can be explained by the simpler switching method, where in LRS the baseline devices have a fully extended filament. Instead, in the $HfO_2 - MO_x$ devices other parts of the material stack are also active.

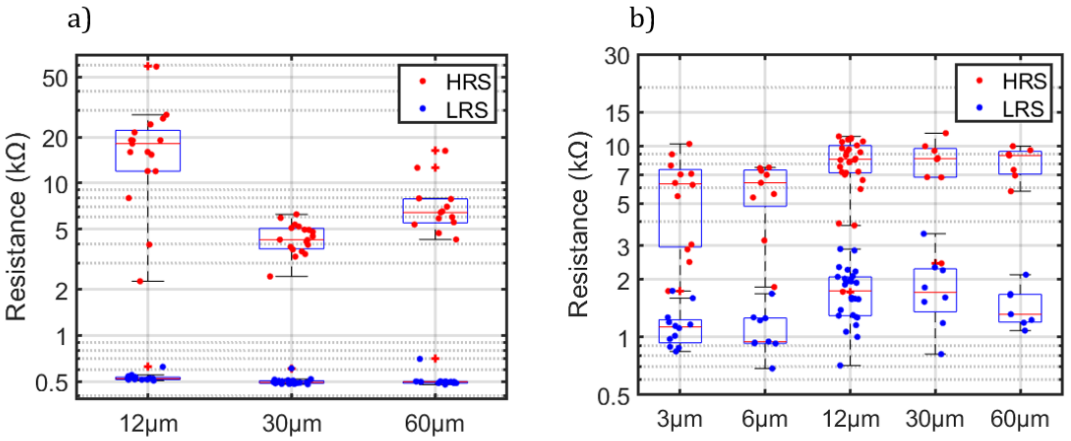

Figure 4.10: Boxplots representing the median HRS and LRS values during cycling, on multiple devices grouped by device size for a) the $HfO_2 - Ti$ technology, and b) the $HfO_2 - MO_x$ technology.

On-Off Ratio The $HfO_2 - MO_x$ technology shows significantly lower On-Off ratio with respect to the baseline technology, Figure 4.11. This parameter reveals the trade-off between the abruptness of the transitions and the On-Off ratio.

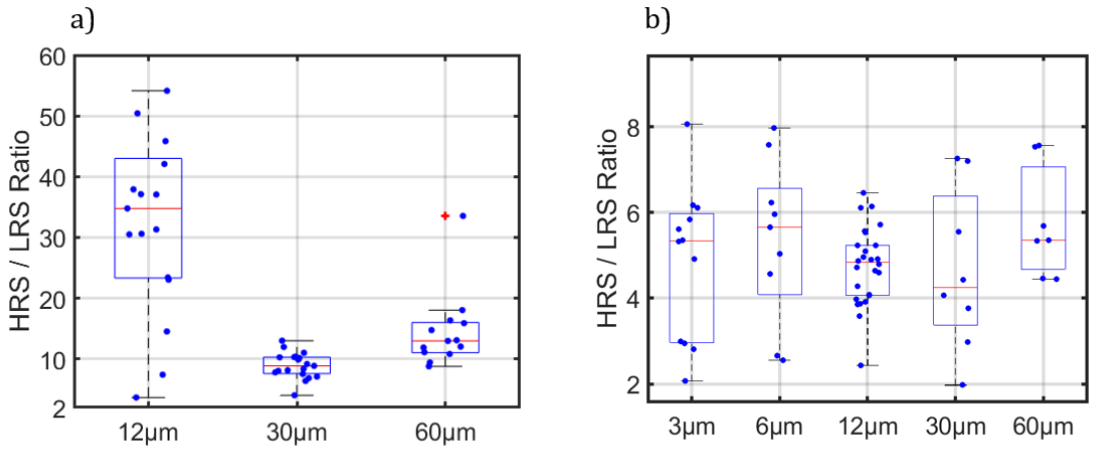

Figure 4.11: Boxplots representing On-Off ratio during cycling, on multiple devices grouped by device size for a) the $HfO_2 - Ti$ technology, and b) the $HfO_2 - MO_x$ technology.

Abruptness of the transitions By calculating the maximum change in resistance within 100 mV during the Set or Reset transition and normalizing it to the HRS value, as detailed in Section 3.1.4, it is possible to quantify the abruptness of the set or reset transition. The lower the relative maximum change in resistance, the more the device is gradually transitioning from the HRS to the LRS or vice versa. A lower abruptness implies that the device can have more intermediate states between the HRS and the LRS.
Confirming the example in Figure 4.9, in Figure 4.12 a it is shown that the baseline devices have a very steep Set transition, while the $HfO_2 - MO_x$ devices are much more analog, Figure 4.12 c. For both devices the Reset transition, Figures 4.12 [b,d], is less abrupt but also in this case the baseline

technology is greatly inferior.
Having such a smooth transition enables the $HfO_2 - MO_x$ devices to be exploited as analog, multilevel memory elements. This is the greatest advantage af the technology with respect to the $HfO_2 - Ti$ devices and it will be more evident for pulsed characterization. As explained in the previous sections, the switching mechanism of the baseline technology relies on self-accelerated (re-)creation and partial destruction cycles of the conductive filament, whereas in $HfO_2 - MO_x$ technology, the hypotheses is that the switching mechanism relies on small changes in the stoichiometry of a small MO_x region at the interface with the filament in the HfO_2.

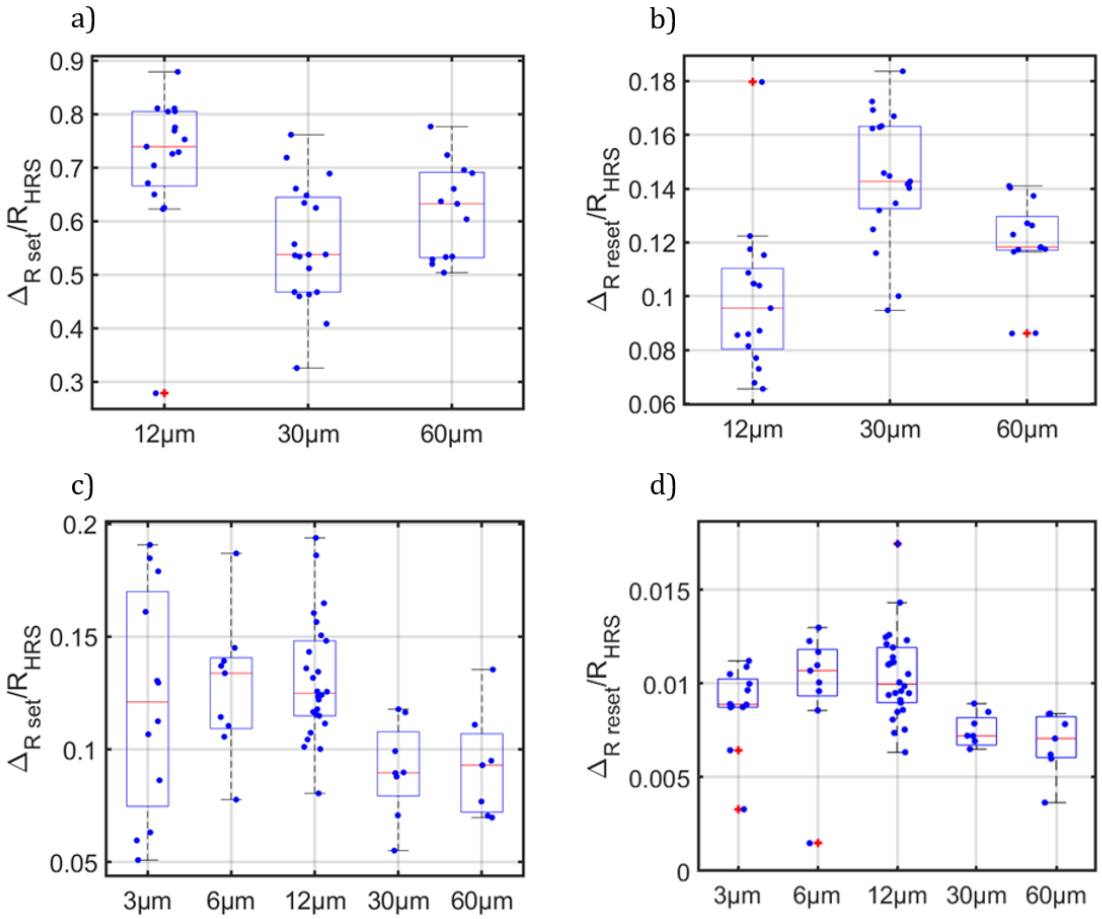

Figure 4.12: Boxplots representing the maximum change in resistance within 100 mV during the Set [a,c] or Reset [b,d] transition and normalizing it to the HRS value, on multiple devices grouped by device size for [a,b] the $HfO_2 - Ti$ technology, and [c,d] the $HfO_2 - MO_x$ technology.

Cycle-to-cycle variability The cycle-to-cycle variability can be quantified by calculating the standard deviation of the HRS and LRS values measured for different cycles, Section 3.1.4. From Figure 4.13 [a,c] it is evident that the baseline devices present a higher cycle-to-cycle variability in the HRS value. This can be related to the fact that during the Reset operation the conductive filament in the HfO_2 is weakened until it is interrupted at a single point. This interruption causes an increase of the resistance. Since the filament is interrupted and reformed each cycle, the spatial and energetic position of the defects that are at the interruption of the filament can change, therefore also the resistance associated to the interruption change, resulting in a high cycle-to-cycle variability. In the $HfO_2 - MO_x$ devices this effect is not present since the switching is dominated by a more oxidized region in the MO_x which does not have such a single point of interruption. Therefore the

technology is less prone to cycle-to-cycle variability.
This hypothesis is supported when observing the standard deviation of the LRS in the two technologies. In the $HfO_2 - MO_x$ devices, the standard deviation of the HRS is comparable to that of the LRS, confirming the hypothesis regarding the more oxidized region in the MO_x, since this is the dominant contribution for both HRS and LRS resistance and thus also for the switching behavior. Additionally, in the $HfO_2 - Ti$ devices the variability of the LRS is extremely low due to the fact that the Set transition is a self-accelerated process and the filament fully forms every time, which leaves little opportunity for variability.

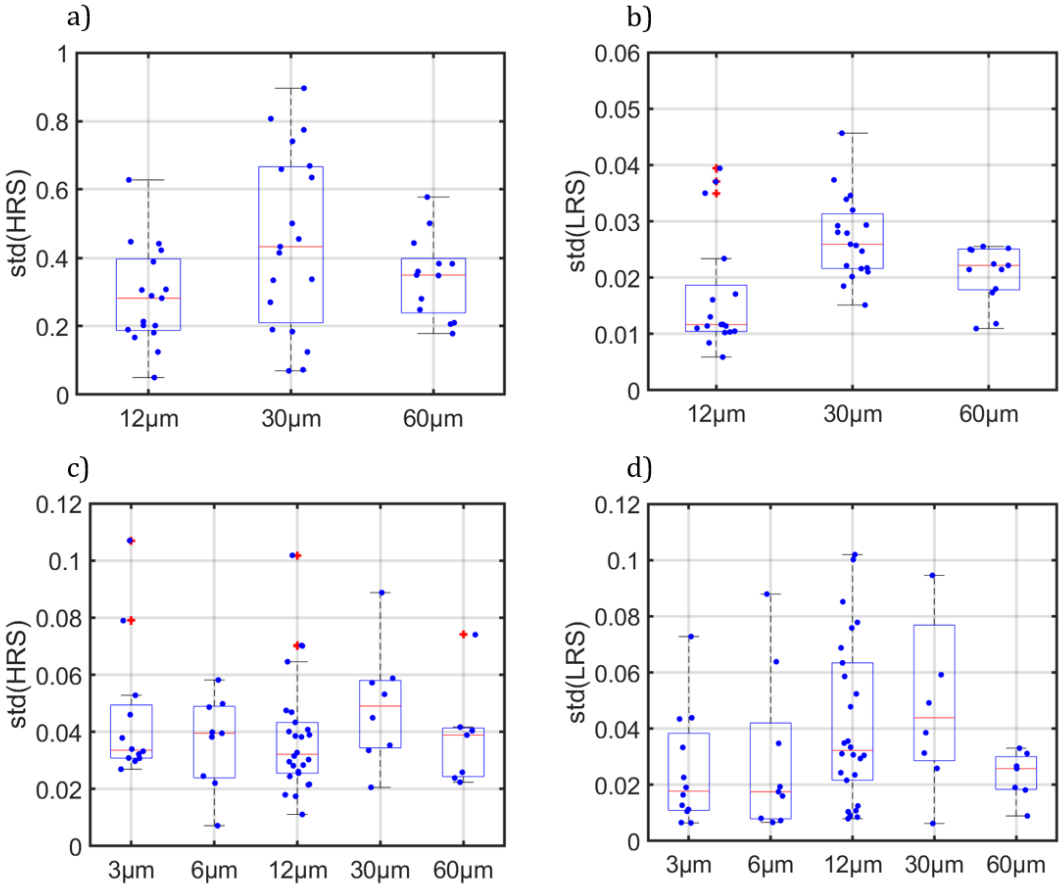

Figure 4.13: Boxplots representing the standard deviation of the HRS [a,c] and LRS [b,d] across the cycle performed, on multiple devices grouped by device size for [a,b] the $HfO_2 - Ti$ technology, and [c,d] the $HfO_2 - MO_x$ technology.

4.1.5 AC Cycling

With the pulsed characterization it is possible to better understand the potential of the two technologies for their final application: memristive elements with the possibility to vary the resistance upon the stimulation with programming pulses.
As illustrated in Section 4.1.4, the $HfO_2 - Ti$ technology present abrupt Set/Reset transitions. This implies that the change in resistance is not gradual as desired. During the AC characterization this is confirmed, as can be seen in Figure 4.14. The device is operated with pulses of 1 µs with amplitudes corresponding to the DC V_{stops}. With these conditions the device switches between the two states, HRS and LRS, in a single pulse, showing no intermediate level and thus a binary operation.
However, if the pulse amplitudes are reduced it is possible to obtain a more analog resistance change

at the expenses of the dynamic range, Figure 4.15. Even though the device shows gradual potentiation and depression curves, the On-Off ratio is too limited for computational applications. If the dynamic range is too limited, the performance of a crossbar array designed with these devices, will not be acceptable for training or inference in DNN applications.
Instead, the novel $HfO_2 - MO_x$ technology offers good On-Off ratio while maintaining a very gradual and repeatable change in resistance by applying the programming pulses, Figure 4.16. A drawback is that switching requires higher pulse amplitudes.

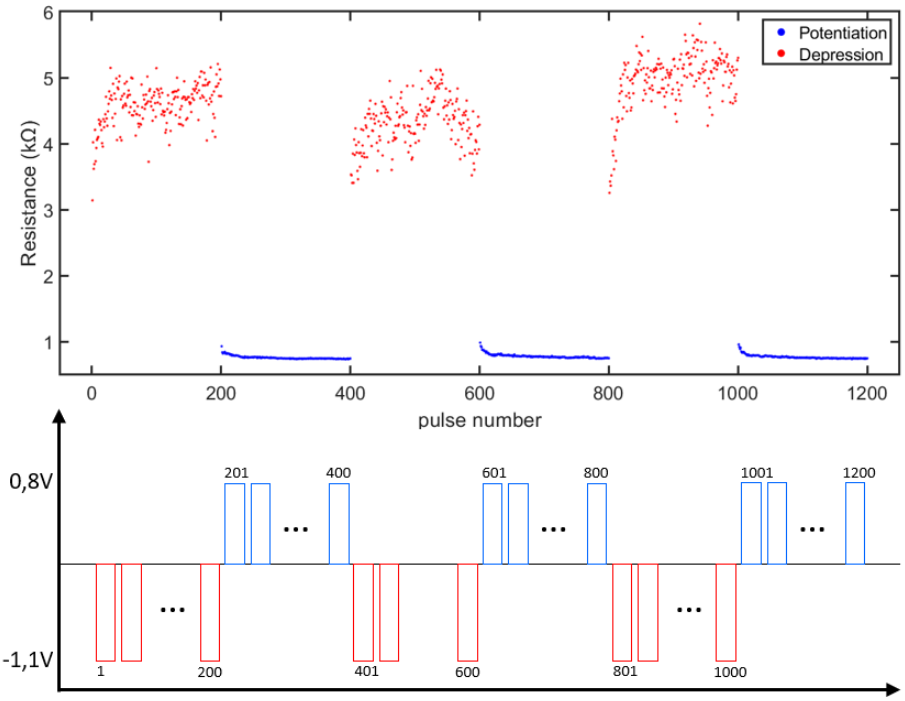

Figure 4.14: Resistance as a function of the applied 1 µs pulses of an $HfO_2 - Ti$ device cycled 3 times. The pulse amplitudes and the applied pulse scheme are visible in the bottom part of the figure.

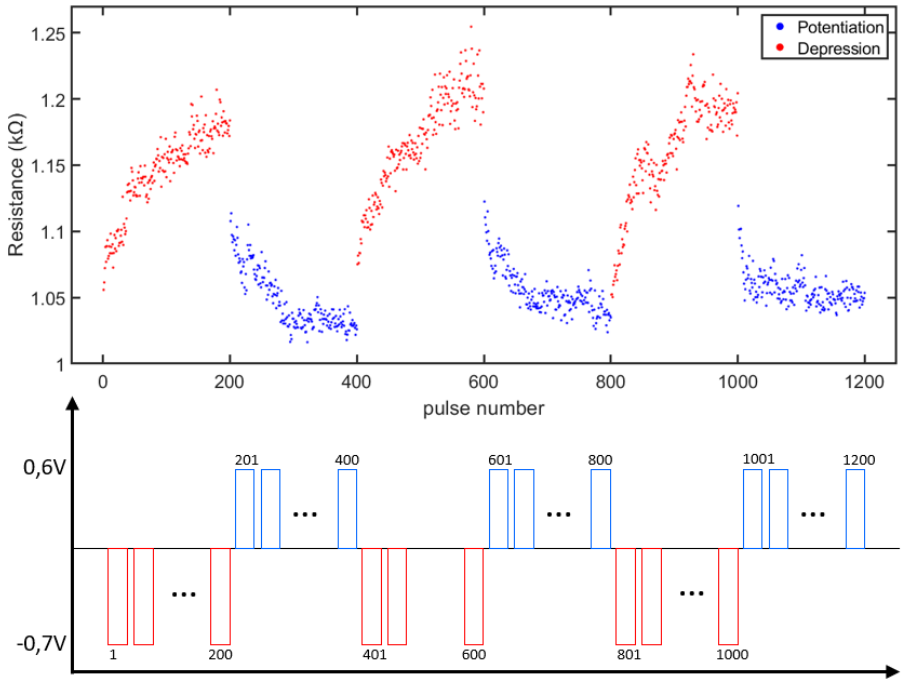

Figure 4.15: Resistance as a function of the applied 1 µs pulses of an $HfO_2 - Ti$ device cycled 3 times. The pulse amplitudes and the applied pulse scheme are visible in the bottom part of the figure.

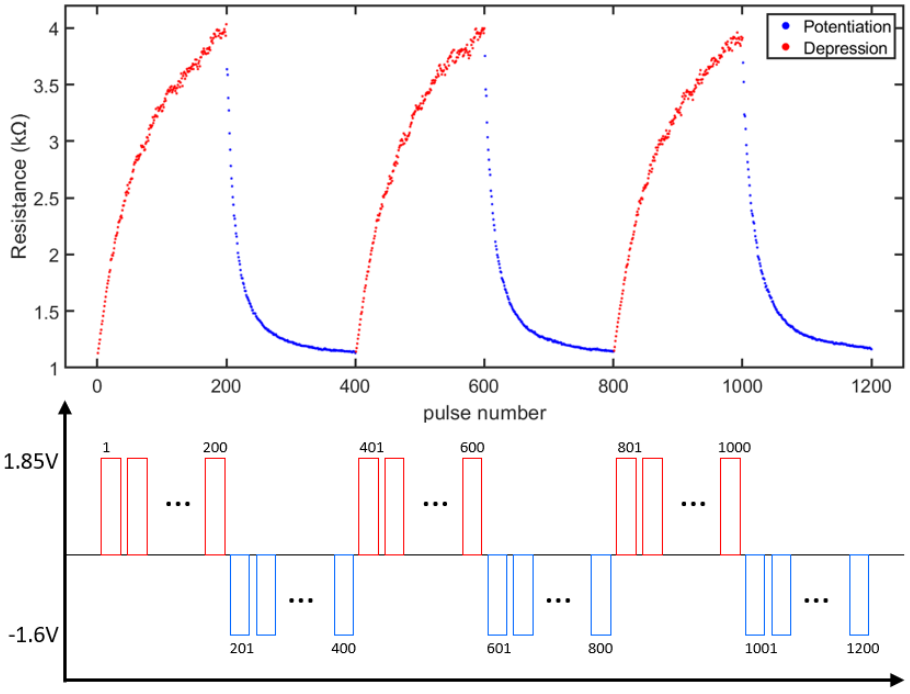

Figure 4.16: Resistance as a function of the applied 1 µs pulses of an $HfO_2 - MOx$ device cycled 3 times. The pulse amplitudes and the applied pulse scheme are visible in the bottom part of the figure.

4.2 Symmetry point Analysis

In Section 4.1.5, it was shown that $HfO_2 - MO_x$ devices can gradually change their conductance. This property can be exploited to perform training of ANNs with the Tiki-Taka algorithm, Section ??. This algorithm relaxes the constraints on the linearity of the potentiation and depression curves. To perform training with the Tiki-Taka algorithm, the devices should have a stable symmetry point. Therefore, it is important to evaluate the reaction of the devices to the symmetrization procedure using the figures of merit described in Section 3.2.3.

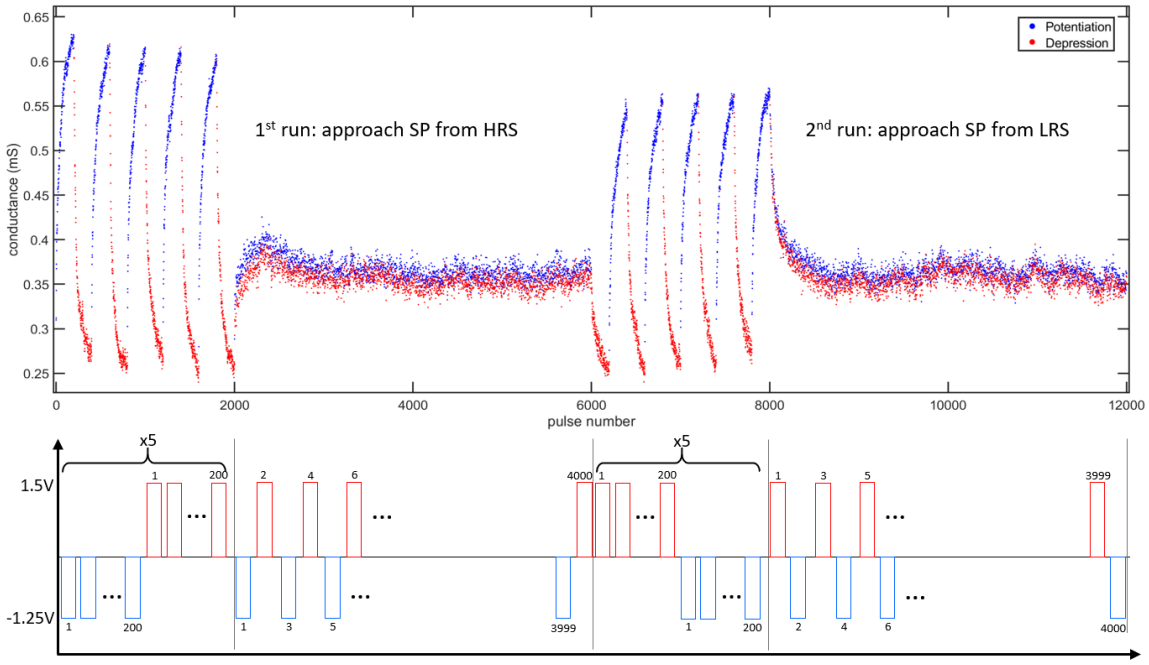

Figure 4.17: Resistance as a function of the applied 600 ns pulses of an $HfO_2 - MOx$ device cycled 5 times and symmetrized for 4000 pulses for two times. The pulse amplitudes and the applied pulse scheme are visible in the bottom part of the figure.

In Figure 4.17 an $HfO_2 - MO_x$ is tested by performing 5 cycles and a 4000 pulses symmetrization procedure twice, the first time the symmetry point is approached from the HRS, while the second time from the LRS. During cycling the device shows a stable analog conductance change with a satisfying On-Off ratio >2.
The device also converges to a symmetry point, which is not perfectly centered in the G_{range} but it is stable and shows a clear separation between the potentiation and depression pulse responses.
However, it must be noted that the dynamic range and the program noise are reduced in the second run which indicates a possible degradation of the performance upon the application of a significant number of pulses.

4.2.1 Performance in IBM AiHwKit

The performances of a crossbar array having as single elements the device in Figure 4.17 can be simulated by using the IBM AiHwKit. For this simulation the device response needs to be modeled. In the IBM AiHwKit, the model which can describe the analyzed device the most accurately is the "PowStepDevice". By fitting the potentiation and depression curves of Figure 4.17, normalized to the interval [-1,1], with the "PowStepDevice", it is possible to find the parameters of this model that best describe the device response, Figure 4.18.

	1st run	2nd run
R_{HRS_AC}	$3.8\,\text{k}\Omega = \frac{1}{262\,\mu\text{S}}$	$3.8\,\text{k}\Omega = \frac{1}{261\,\mu\text{S}}$
R_{LRS_AC}	$1.6\,\text{k}\Omega = \frac{1}{612\,\mu\text{S}}$	$1.8\,\text{k}\Omega = \frac{1}{556\,\mu\text{S}}$
$On - Off_{ratio}$	2.33	2.13
G_{range}	$350\,\mu\text{S}$	$295\,\mu\text{S}$
R_{SP}	$2.8\,\text{k}\Omega = \frac{1}{356\,\mu\text{S}}$	$2.8\,\text{k}\Omega = \frac{1}{359\,\mu\text{S}}$
ΔG_{SP}^{mean}	$10.5\,\mu\text{S}$	$7.5\,\mu\text{S}$
SP_{skew_HRS}	27%	33%
Program noise	98%	128%
#states	33	40

Table 1: Table showing the figures of merit extracted from the measure in Figure 4.17.

Using the extracted parameters the training accuracy of a crossbar array implementing a single fully connected layer on the MNIST dataset can be simulated. The ANN implemented by the crossbar array consisting of the modeled $HfO_2 - MO_x$ devices has an accuracy of 81%, while the accuracy of the ANN with FP64 precision is 90%.

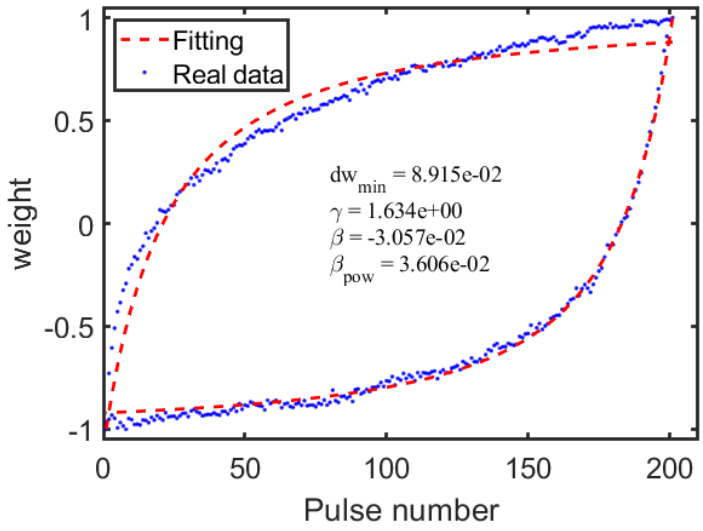

Figure 4.18: Potentiation and depression curves of Figure 4.17 normalized to the interval [-1,1], as a function of the number of pulses and the fitting with the "PowStepDevice" of IBM AiHwKit.

4.3 Temperature-dependent DC measurements

This section reports the results of several temperature-dependent experiments on $HfO_2 - Ti$ and $HfO_2 - MOx$ devices. The goal of these experiments is to gain more information about the physical principles of the operation of the analyzed RRAM technologies. In the following sections the following experiments will be reported: Pristine sweeps, Forming, Cyling at room temperature after temperature-dependent forming, "hot forming", HRS and LRS resistances and cycling after forming at room temperature.

4.3.1 Pristine sweeps: $HfO_2 - Ti$ technology

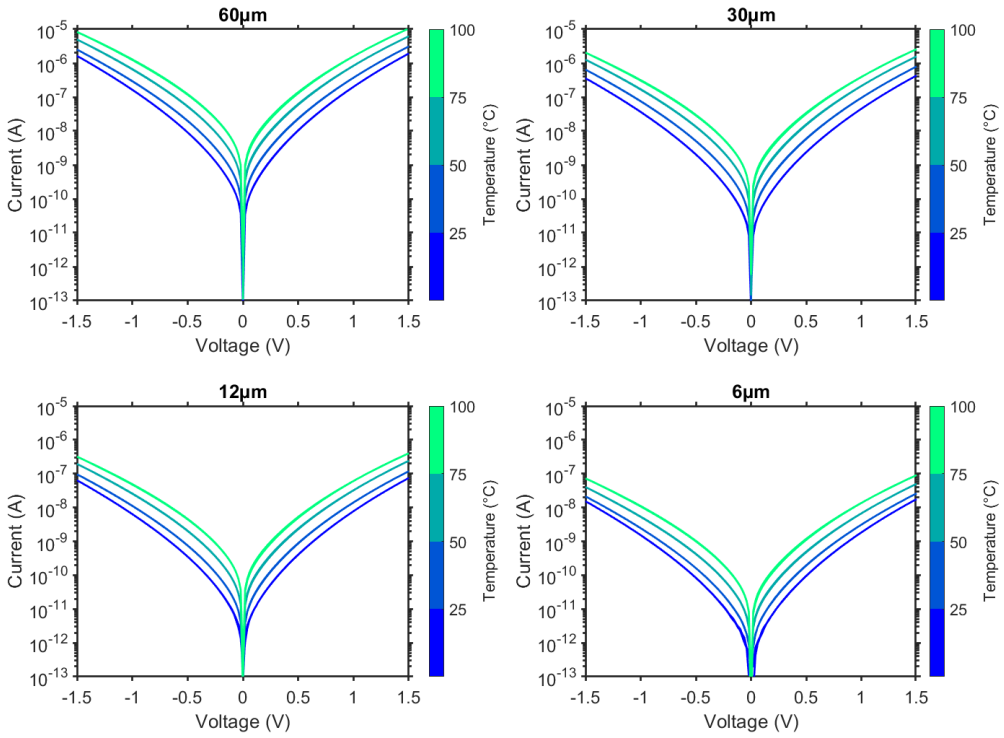

Figure 4.19: IV, current-voltage, curves in pristine state of $HfO_2 - Ti$ devices, for different sizes measured at different temperatures. The plotted IV curves are from 0 V to 1.5 V to −1.5 V to 0 V.

In Figure 4.19 the leakage current for the $HfO_2 - Ti$ technology is shown for different sizes and temperatures. Each curve represents a different device, never previously measured, in order to avoid any effects due to previous measurements. The consistency of the results is high since the devices are in pristine state, meaning the material stack has not been measured after processing.

From Figure 4.19 it can be observed that the leakage current increases with increasing device size and with increasing temperature. These trends are expected, since a larger device offers more conduction paths through the dielectric layer, while a higher temperatures favours the tunneling processes that are responsible for the conduction in the HfO_2 layer.
The baseline devices do not show any hysteresis and no difference if the pristine sweep is performed in the opposite direction: from 0 V to −1.5 V to 1.5 V to 0 V. Moreover, the pristine sweeps in Figure 4.19 are symmetric, indicating that the conduction for positive and negative voltages is similar, that for the most part, it can be described by bulk-controlled phenomena and that there are no major interface effects. Although, it has to be noted that the leakage current for positive biases is slightly higher with respect to the negative biases, suggesting a small asymmetry in the interfaces of the

HfO_2 with the top and bottom electrodes, which is reasonable considering that they are two different materials TiN and TiOx – Ti.

In Figure 4.20 the JV curves for the various sizes overlap for all the temperatures. This indicates

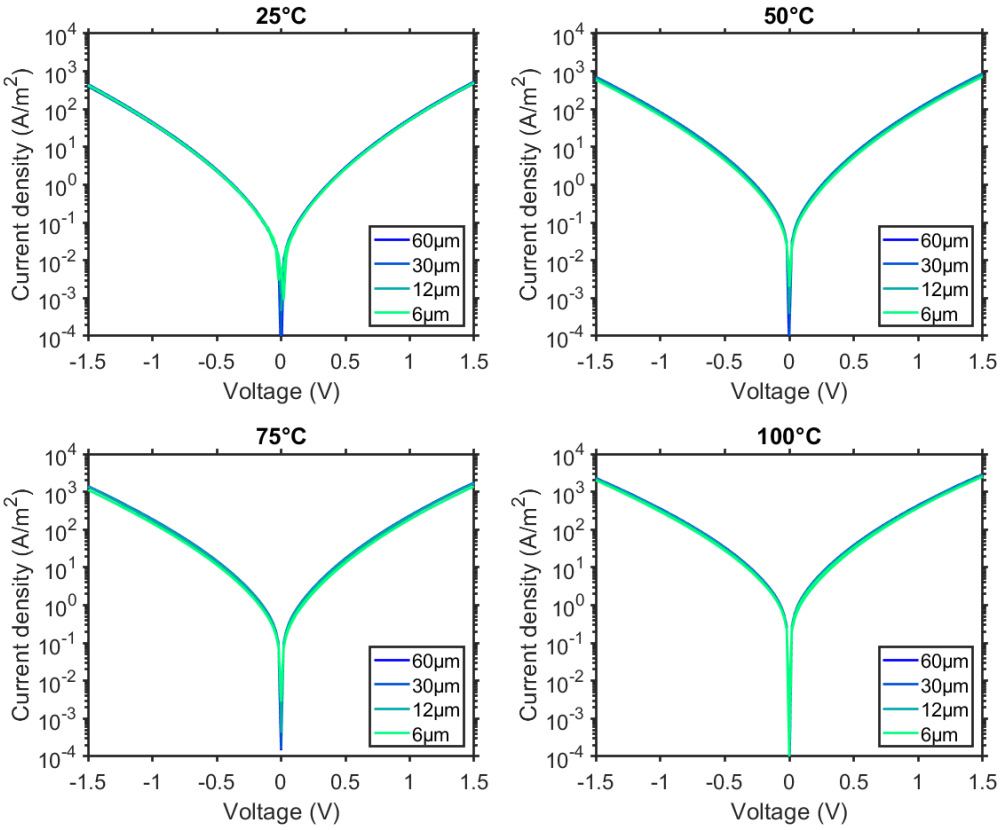

Figure 4.20: J-V, current density-voltage, curves in pristine state of HfO_2 – Ti devices, for different sizes measured at different temperatures. The plotted IV curves are from 0 V to 1.5 V to −1.5 V to 0 V.

an exact area dependence of the leakage current, which implies that the conduction in pristine state in the baseline technology takes place through the whole device area and that it has no preferential path.

Conduction mechanisms fitting By fitting the J-V curve, or a manipulation such as $\log(J/V) - \sqrt{V}$, with a straight line, it is possible to deduce physical parameters from the slope and the intercept of the straight using the equations describing the conduction mechanisms reported in Section 3.3.3. If a conduction mechanism can explain the conduction in the fitted region, the extracted physical parameters should be reasonable for the analyzed material, HfO_2. If not, the extracted parameters will be far from the values that can be found in literature for the material. However the models are extremely simple and cannot fully represent the actual physical operation of the device, but they can be a useful tool if it is taken into account that they are just a first approximation. For a more complete modeling the Ginestra software, or a similar environment, needs to be used.

Ohmic Conducion fitting Fitting the first four data points of the pristine sweep, from 0.02 V to 0.08 V, Figure 4.21, enables the extraction of the parameters listed in the table in

Figure 4.22 by using the formula for the Ohmic conduction, Equation 3.24. It is reasonable to have Ohmic conduction at very low biases; when the electric field is low there is not enough energy for any other conduction mechanism to dominate, therefore the conduction is due to drift in the bands of the intrinsic carriers on the tail of the energy distribution, which have enough energy to tunnel to the bands.
The extracted values listed in the table in Figure 4.22 are in good agreement with each other, except for the 6 µm devices with the pristine sweep performed initially at positive biases. This is probably an outlier, indicating a slight variability of the devices in pristine state.
The fit has a slope of 1.1 in the $\log J - \log V$ plane, close to 1 which is the expected slope for Ohmic conduction, and $\mu \cdot N_c$, the product of the mobility and the number of states in the Conduction Band, is reasonable for an insulator such as HfO_2. However, the difference between the Fermi level and the bottom of the Conduction Band is incredibly small for HfO_2, which has a bandgap between 5.5 eV and 6 eV. Thus, it is possible that this value represents other energy differences such as the difference between the work functions of the electrodes and the bottom of the Conduction Band of the HfO_2 or the trap level in the HfO_2 or, more probably, it indicates that this model cannot fully describe the conduction in the device.

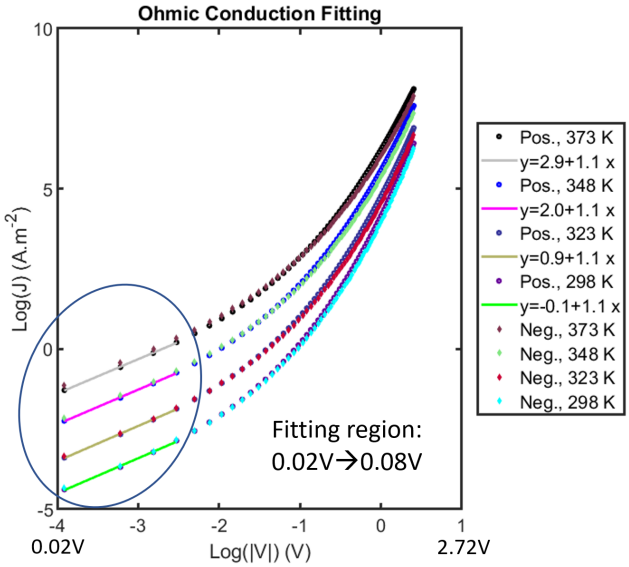

Figure 4.21: $\log J - \log V$ graphs with fittings of the pristine sweeps performed at different temperatures on $HfO_2 - Ti$ devices.

0V → 1.5V, 0V → -1.5V

Size	Pos $E_C - E_F$	Neg $E_C - E_F$	Pos $\mu \cdot N_c$	Neg $\mu \cdot N_c$
60 um	0.40 eV	0.41 eV	1.19e+15 $\frac{1}{cm \cdot V \cdot s}$	2.01e+15 $\frac{1}{cm \cdot V \cdot s}$
30 um	0.40 eV	0.42 eV	1.67e+15 $\frac{1}{cm \cdot V \cdot s}$	2.92e+15 $\frac{1}{cm \cdot V \cdot s}$
12 um	0.40 eV	0.40 eV	2.13e+15 $\frac{1}{cm \cdot V \cdot s}$	2.64e+15 $\frac{1}{cm \cdot V \cdot s}$
6 um	0.58 eV	0.55 eV	5.22e+18 $\frac{1}{cm \cdot V \cdot s}$	2.25e+18 $\frac{1}{cm \cdot V \cdot s}$

0V → -1.5V, 0V → 1.5V

Size	Pos $E_C - E_F$	Neg $E_C - E_F$	Pos $\mu \cdot N_c$	Neg $\mu \cdot N_c$
60 um	0.38 eV	0.38 eV	6.22e+14 $\frac{1}{cm \cdot V \cdot s}$	5.10e+14 $\frac{1}{cm \cdot V \cdot s}$
30 um	0.41 eV	0.41 eV	2.39e+15 $\frac{1}{cm \cdot V \cdot s}$	2.31e+15 $\frac{1}{cm \cdot V \cdot s}$
12 um	0.40 eV	0.39 eV	2.27e+15 $\frac{1}{cm \cdot V \cdot s}$	1.66e+15 $\frac{1}{cm \cdot V \cdot s}$
6 um	0.39 eV	0.37 eV	7.01e+15 $\frac{1}{cm \cdot V \cdot s}$	3.42e+15 $\frac{1}{cm \cdot V \cdot s}$

Figure 4.22: Tables showing the parameters extracted from the Ohmic fitting in Figure 4.21 using Equation 3.24. Pos and Neg show the parameters extracted at either positive or negative biases.

Poole-Frenkel Emission fitting At high biases, from 0.7 V to 1.5 V the Poole-Frenkel mechanism can well describe the leakage current in the analyzed stack as shown in Figure 4.23. Using Equation 3.22, physical parameters, such as the distance between the trap energy level and the bottom of the Conduction Band of the HfO_2, Φ_T, the relative dielectric constant ϵ_r and the product of the mobility and the number of states in the Conduction Band, $\mu \cdot N_c$, can be extracted and are reported in the table in Figure 4.24.
As for the Ohmic conduction, the extracted values are very similar apart from the 6 µm devices with positive pristine sweep first, indicating a slight variability in pristine state.
The $\mu \cdot N_c$ values are similar to those extracted by the Ohmic conduction fit. This gives a self-confirmation of the two fittings since they predict the same material parameters.
The extracted relative dielectric constant is close to 21, which is a common value found in literature for HfO_2 [30],[35], however it is debated whether the ϵ_r extracted by the Poole-Frenkel fit is the static or the dynamic (optical) one, which is close to 2, [19].
The extracted Φ_T, around 0.5 eV, would suggest that the traps in the HfO_2 are extremely close to the bottom of the Conduction Band, which is quite unexpected, but it must be taken into account that these models are an extreme simplification of the real processes happening in the device. Considering this fact, as proposed for the Ohmic conduction, it is also possible that this energy difference is an "effective" Φ_T which masks some more complex phenomena not described within this particular model, such as the injection and extraction of carriers into the HfO_2 from the electrodes and the tunneling trough the traps present in the dielectric layer.

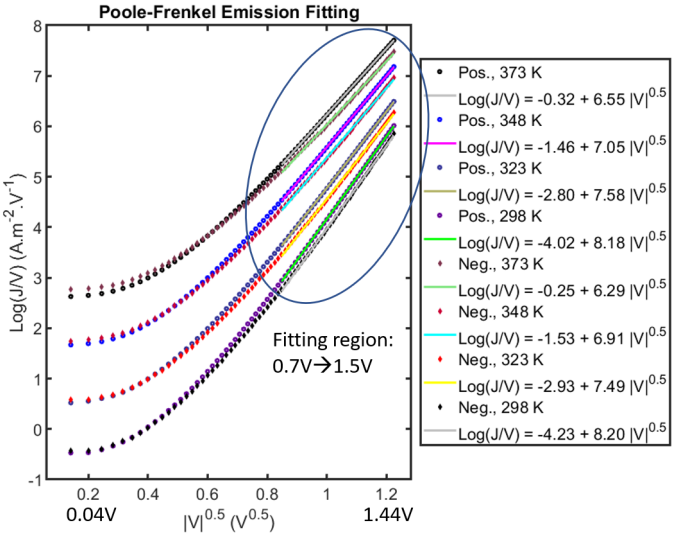

Figure 4.23: $\log(J/V) - \sqrt{V}$ graphs with fittings of the pristine sweeps performed at different temperatures on HfO_2 – Ti devices.

0V → 1.5V, 0V → -1.5V

Size	Pos ϕ_T	Neg ϕ_T	Pos ε_r	Neg ε_r	Pos $\mu \cdot N_c$	Neg $\mu \cdot N_c$
60 um	0.48 eV	0.51 eV	22.3	23.1	6.58e+14 $\frac{1}{cm \cdot V \cdot s}$	1.95e+15 $\frac{1}{cm \cdot V \cdot s}$
30 um	0.48 eV	0.52 eV	22.3	23.1	8.93e+14 $\frac{1}{cm \cdot V \cdot s}$	2.84e+15 $\frac{1}{cm \cdot V \cdot s}$
12 um	0.47 eV	0.51 eV	22.2	22.9	1.11e+15 $\frac{1}{cm \cdot V \cdot s}$	3.13e+15 $\frac{1}{cm \cdot V \cdot s}$
6 um	0.63 eV	0.66 eV	22.2	22.9	1.39e+18 $\frac{1}{cm \cdot V \cdot s}$	3.81e+18 $\frac{1}{cm \cdot V \cdot s}$

0V → -1.5V, 0V → 1.5V

Size	Pos ϕ_T	Neg ϕ_T	Pos ε_r	Neg ε_r	Pos $\mu \cdot N_c$	Neg $\mu \cdot N_c$
60 um	0.46 eV	0.47 eV	22.4	22.6	3.67e+14 $\frac{1}{cm \cdot V \cdot s}$	4.05e+14 $\frac{1}{cm \cdot V \cdot s}$
30 um	0.49 eV	0.49 eV	22.4	22.5	1.20e+15 $\frac{1}{cm \cdot V \cdot s}$	1.13e+15 $\frac{1}{cm \cdot V \cdot s}$
12 um	0.47 eV	0.48 eV	22.3	22.4	1.16e+15 $\frac{1}{cm \cdot V \cdot s}$	1.05e+15 $\frac{1}{cm \cdot V \cdot s}$
6 um	0.46 eV	0.47 eV	22.3	22.2	2.98e+15 $\frac{1}{cm \cdot V \cdot s}$	2.72e+15 $\frac{1}{cm \cdot V \cdot s}$

Figure 4.24: Tables showing the parameters extracted from the Poole-Frenkel fitting in Figure 4.23 using Equation 3.22. Pos and Neg show the parameters extracted at either positive or negative biases.

Hopping Conduction fitting In the $\log(J) - V$ plot, the pristine sweep can be fitted with a straight line from 0.9 V to 1.5 V, Figure 4.25, indicating that the conduction can be explained with the Hopping model, Section 3.3.3. Using Equation 3.23 it is possible to calculate: the distance between the trap energy level and the bottom of the Conduction Band of the HfO_2, Φ_T, the mean Hopping distance, a and the product between the electron concentration in the conduction band of the dielectric and the frequency of thermal vibration of the electrons at trap sites, $\nu \cdot n$. The extracted values are reported in the table in Figure 4.26 and as for the previous fitting present a slight variability.
A Hopping distance of 7 Å is reasonable, since the thickness of the HfO_2 layer is 5.5 nm. Considering an effective vibration frequency of the O-Hf bonds, $\nu = 7 \times 10^{13}$ Hz, it is possible to calculate the electron density $n \approx 1 \times 10^{14}\,cm^{-3}$, which is coherently lower than the extracted density of states from the previous fits. The extracted energy level of the traps is too shallow to be reasonable and the further considerations made for the Ohmic and the Poole-Frenkel fit are valid also in this case.

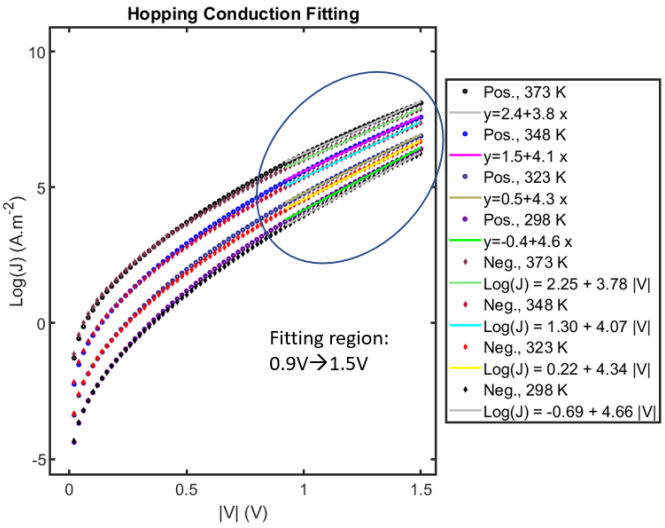

Figure 4.25: $\log(J) - V$ graphs with fittings of the pristine sweeps performed at different temperatures on $HfO_2 - Ti$ devices.

0V → 1.5V, 0V → -1.5V

Size	Pos $E_C - E_T$	Neg $E_C - E_T$	Pos d_H	Neg d_H	Pos $\nu \cdot n$	Neg $\nu \cdot N$
60 um	0.36 eV	0.38 eV	7.0 Å	7.0 Å	7.69e+33 $\frac{1}{m^3 \cdot s}$	9.96e+33 $\frac{1}{m^3 \cdot s}$
30 um	0.37 eV	0.39 eV	7.0 Å	7.0 Å	1.14e+34 $\frac{1}{m^3 \cdot s}$	1.57e+34 $\frac{1}{m^3 \cdot s}$
12 um	0.36 eV	0.38 eV	7.0 Å	7.0 Å	1.33e+34 $\frac{1}{m^3 \cdot s}$	1.70e+34 $\frac{1}{m^3 \cdot s}$
6 um	0.52 eV	0.53 eV	7.1 Å	7.0 Å	1.57e+37 $\frac{1}{m^3 \cdot s}$	1.94e+37 $\frac{1}{m^3 \cdot s}$

0V → -1.5V, 0V → 1.5V

Size	Pos $E_C - E_T$	Neg $E_C - E_T$	Pos d_H	Neg d_H	Pos $\nu \cdot n$	Neg $\nu \cdot N$
60 um	0.35 eV	0.34 eV	7.0 Å	7.1 Å	3.83e+33 $\frac{1}{m^3 \cdot s}$	2.76e+33 $\frac{1}{m^3 \cdot s}$
30 um	0.38 eV	0.38 eV	7.0 Å	7.1 Å	1.41e+34 $\frac{1}{m^3 \cdot s}$	1.01e+34 $\frac{1}{m^3 \cdot s}$
12 um	0.36 eV	0.36 eV	7.0 Å	7.1 Å	1.34e+34 $\frac{1}{m^3 \cdot s}$	8.59e+33 $\frac{1}{m^3 \cdot s}$
6 um	0.35 eV	0.35 eV	7.0 Å	7.1 Å	3.45e+34 $\frac{1}{m^3 \cdot s}$	2.15e+34 $\frac{1}{m^3 \cdot s}$

Figure 4.26: Tables showing the parameters extracted from the Hopping conduction fitting in Figure 4.25 using Equation 3.23. Pos and Neg show the parameters extracted at either positive or negative biases.

4.3.2 Pristine sweeps: HfO_2 – MOx technology

The pristine I-V curves of HfO_2 – MOx devices are reported in Figure 4.27. Each line is the composition of two measurements, from 0 V to 3 V and from 0 V to −3 V. The two measurements are performed on different and never before measured devices. The HfO_2 – MOx technology shows hysteresis and a different behavior if the pristine sweeps are performed in the opposite direction, first negative voltages than positive voltages as detailed in Figure **??**. During the measurements an integration time of 0.1 sec has been used in order to measure lower currents with higher accuracy. Nonetheless, in Figures [4.27,4.28,4.29] the region is highlighted where the current drops below 1×10^{-11} A, which is an extremely conservative lower bound. This highlighting has been done to provide a reference above which the measured current is without doubt unaffected by noise or systematic errors.
The leakage current increases with size and temperature, as expected, also for this technology. However, it is possible to note a strong asymmetric behavior between the positive and the negative biases. At low biases and low temperatures the current is higher for positive voltages, but above 100 °C the concavity of the current for negative voltages changes and the leakage current becomes higher than at positive voltage. At higher voltages, the conduction is enhanced for positive biases, moreover at 150 °C there is a sudden change in the slope of the I-V curve.

From the J-V plots, Figure 4.28, it is clear that the leakage current does not scale with the device area for all voltages. This finding suggests that at low voltages the conduction is predominantly trough a preferential path, such as a filament, and not through the whole surface, while at higher voltages the bulk conduction becomes dominant.
Since the pristine sweeps on the HfO_2 – Ti devices showed a perfect scaling, the preferential path is probably in the MOx layer. This hypothesis is supported by the plots in Figure 4.29, where it is shown that from −1 V to 1 V the current is equal for all the device sizes, implying a similar conduction mechanism, independent of the temperature at which the measurement is performed. Since at high temperatures the currents measured between −1 V to 1 V are mostly higher than 1×10^{-11} A it is certain that this effect is not an artifact created by errors in the measurements, but it is the actual device behavior.

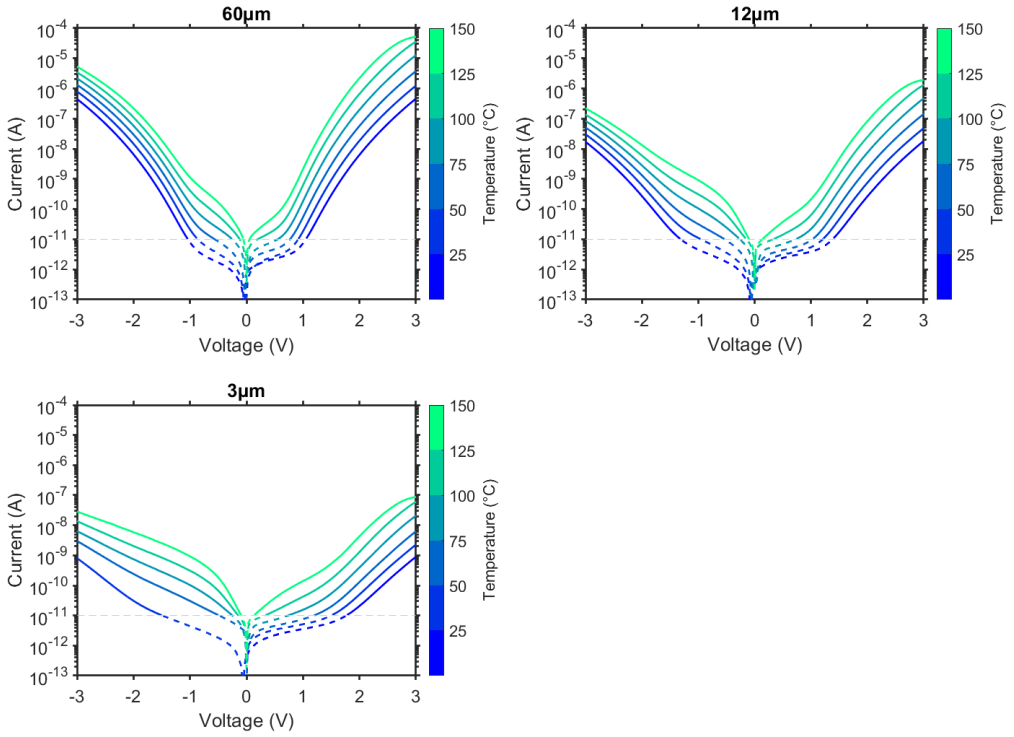

Figure 4.27: I-V, current-voltage, curves in pristine state of $HfO_2 - MOx$ devices, for different sizes measured at different temperatures. The plotted IV curves are from 0 V to 3 V and from 0 V to −3 V. The I-V curve is dashed for currents below 1×10^{-11} A.

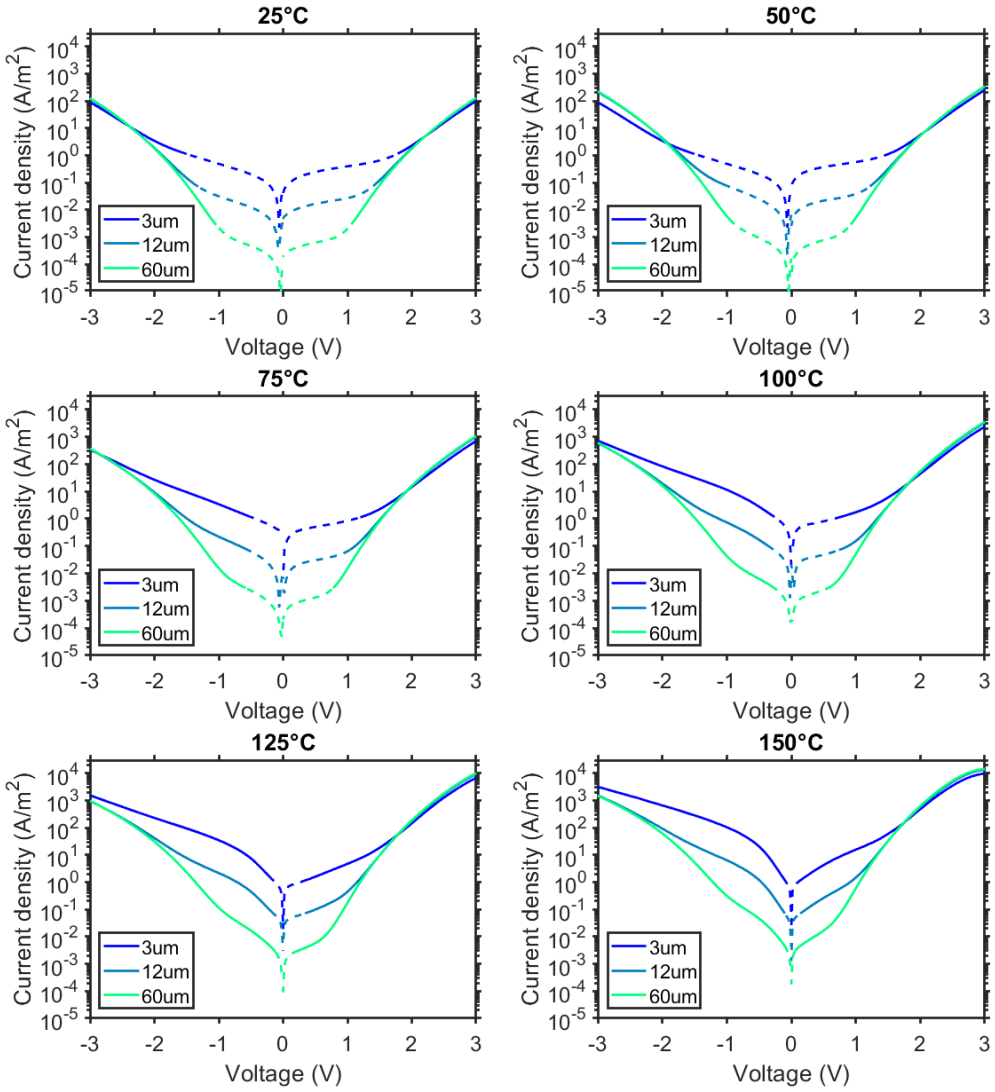

Figure 4.28: J-V, current density-voltage, curves in pristine state of $HfO_2 - MOx$ devices, for different sizes measured at different temperatures. The plotted IV curves are from 0 V to 3 V and from 0 V to −3 V. The J-V curve is dashed for currents below 1×10^{-11} A.

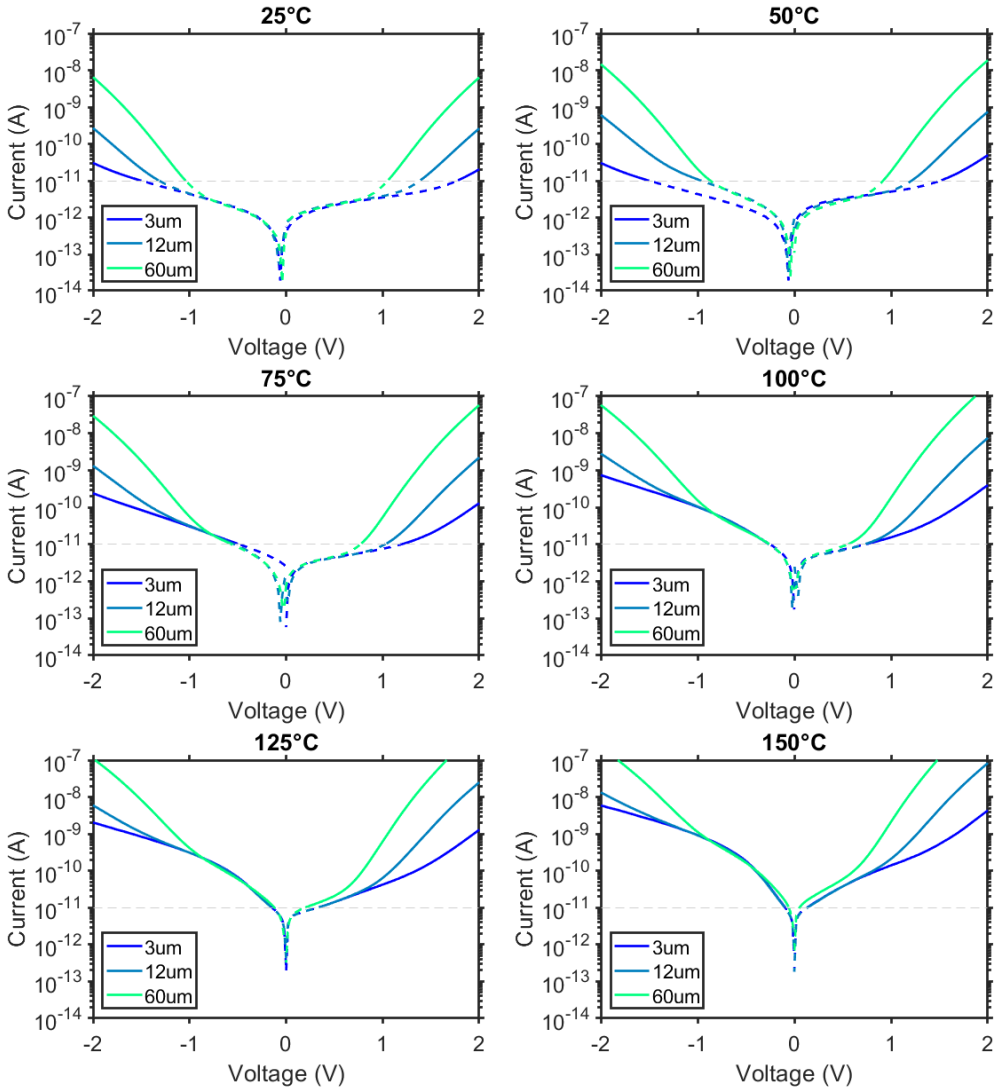

Figure 4.29: I-V, current-voltage, curves in pristine state of $HfO_2 - MOx$ devices, for different sizes measured at different temperatures. The plotted IV curves are from 0 V to 3 V and from 0 V to −3 V. The I-V curve is dashed for currents below 1×10^{-11} A.

Comparison with HfO_2 – Ti technology From the comparison between the pristine sweeps in Figure 4.30, two striking difference are immediately visible: HfO_2 – MOx devices show hysteresis and a leakage current of 2-3 orders of magnitude lower at 1.5 V. These difference can be attributed to the MOx layer, which is a more insulating material with respect to Ti. Firthermore, it can cause an oxidation of the HfO_2 at the interface, Section 4.1.1, and generation of defects can happen in this layer.
From Figure 4.30 a),c) it is possible to note that the hysteresis is enhanced by increasing the temperature and that it is activated only by the positive sweeps, because in Figure 4.30 c) where the negative sweep is performed prior to the positive sweep there is no hysteresis at negative biases. This fact highlights that defects are generated at positive voltages and then moved in the subsequent negative sweep, Figure 4.30 a).

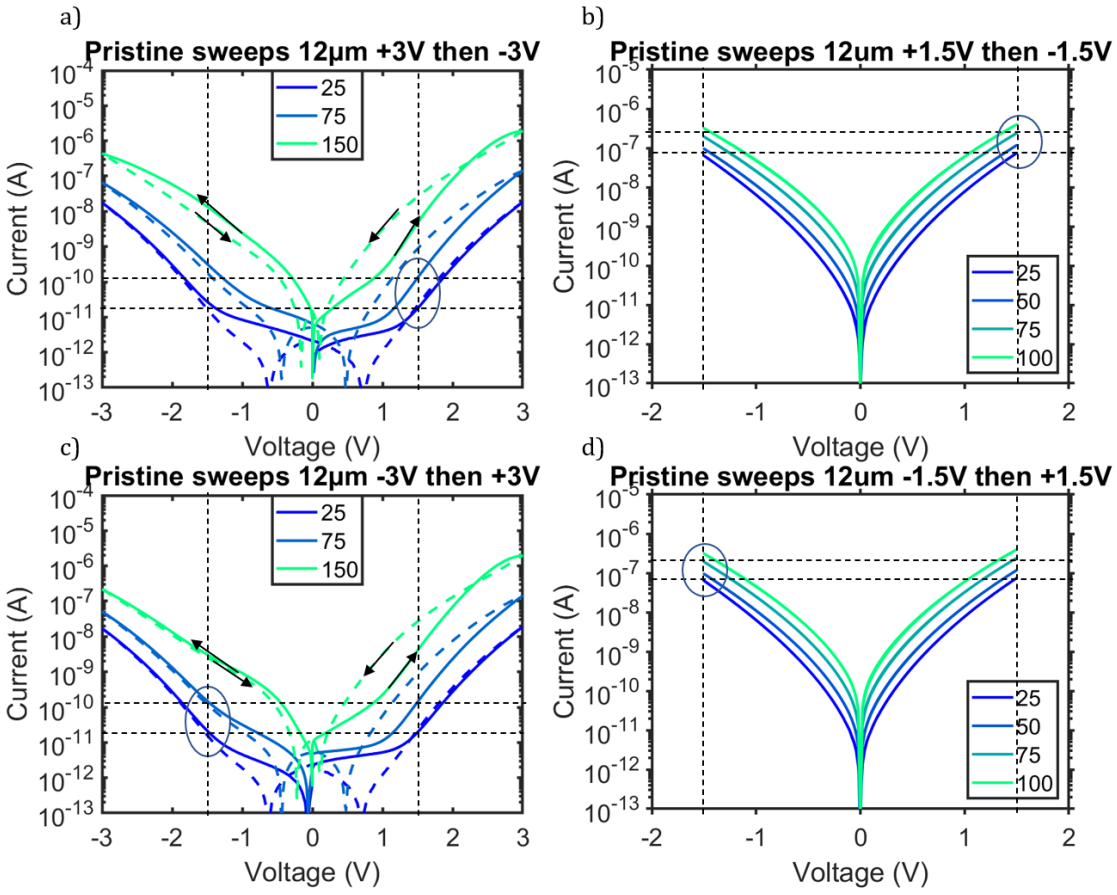

Figure 4.30: I-V, current-voltage, curves in pristine state of HfO_2 – MOx devices a),c) and HfO_2 – Ti devices b),d), for 12 µm devices measured at different temperatures. The plotted IV curves are from 0 V to V_{stop} to $-V_{stop}$ a),b) and from 0 V to $-V_{stop}$ to V_{stop} c),d). For HfO_2 – MOx devices the backward parts of the sweeps, from $\pm V_{stop}$ to 0 V is dashed to highlight the hysteresis. The black dashed lines help to identify the current at 1.5 V for 25 °C and 75 °C.

4.3.3 Forming: HfO_2 – Ti technology

Forming the devices at different temperatures can give insights into the forming process. In Figures 4.31 and 4.32, the forming voltage and current of HfO_2 – Ti devices as a function of the temperature or $1/k_BT$ is shown. For both 12 µm and 60 µm devices the forming voltage decreases with increasing

temperature, indicating that it is a temperature-dependent phenomenon. From Figure 4.32, it is clear that the forming current is not affected by the forming temperature, confirming that the forming is a field-driven process.

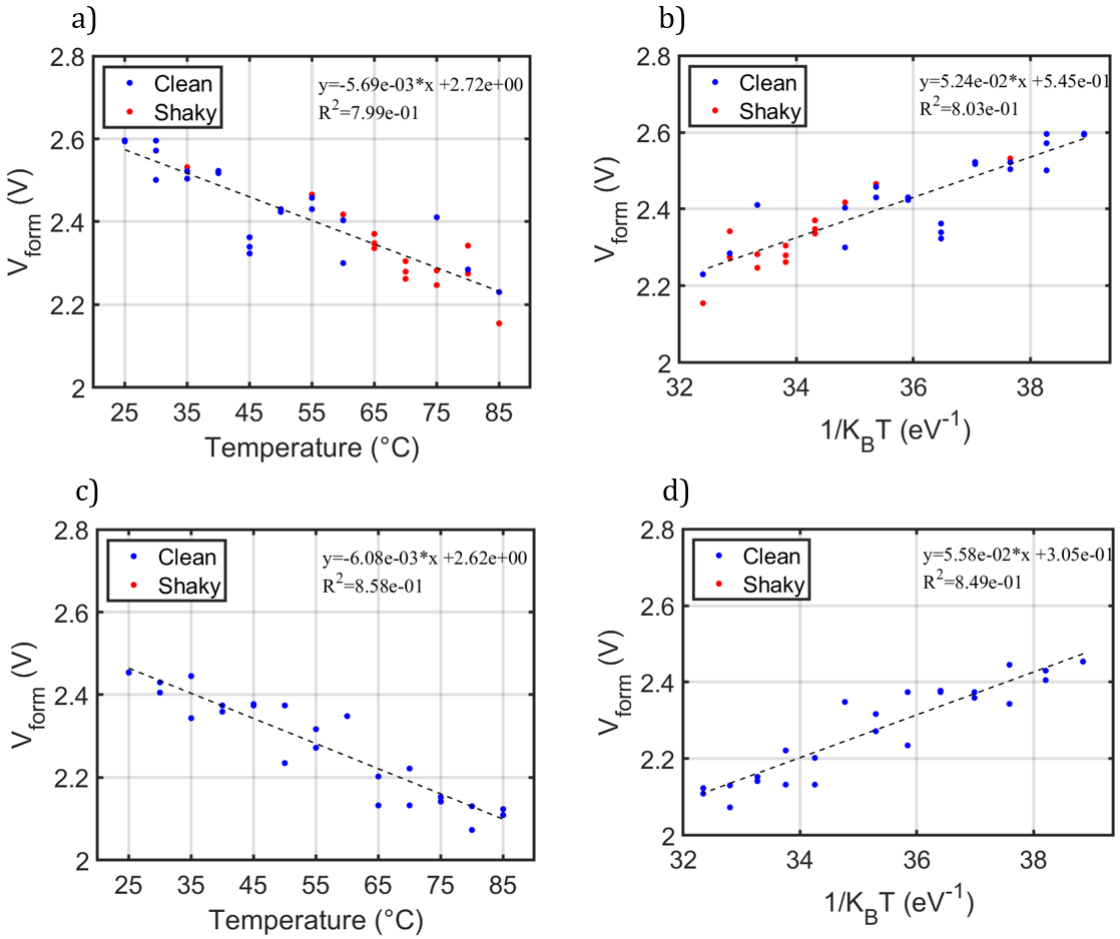

Figure 4.31: Forming voltage, V_{form}, as a function of the forming temperature and $1/k_BT$ for $HfO_2 - Ti$ devices of 12 µm size a),b), and 12 µm size c),d). The red dots highlight a "shaky", not in a single step, forming.

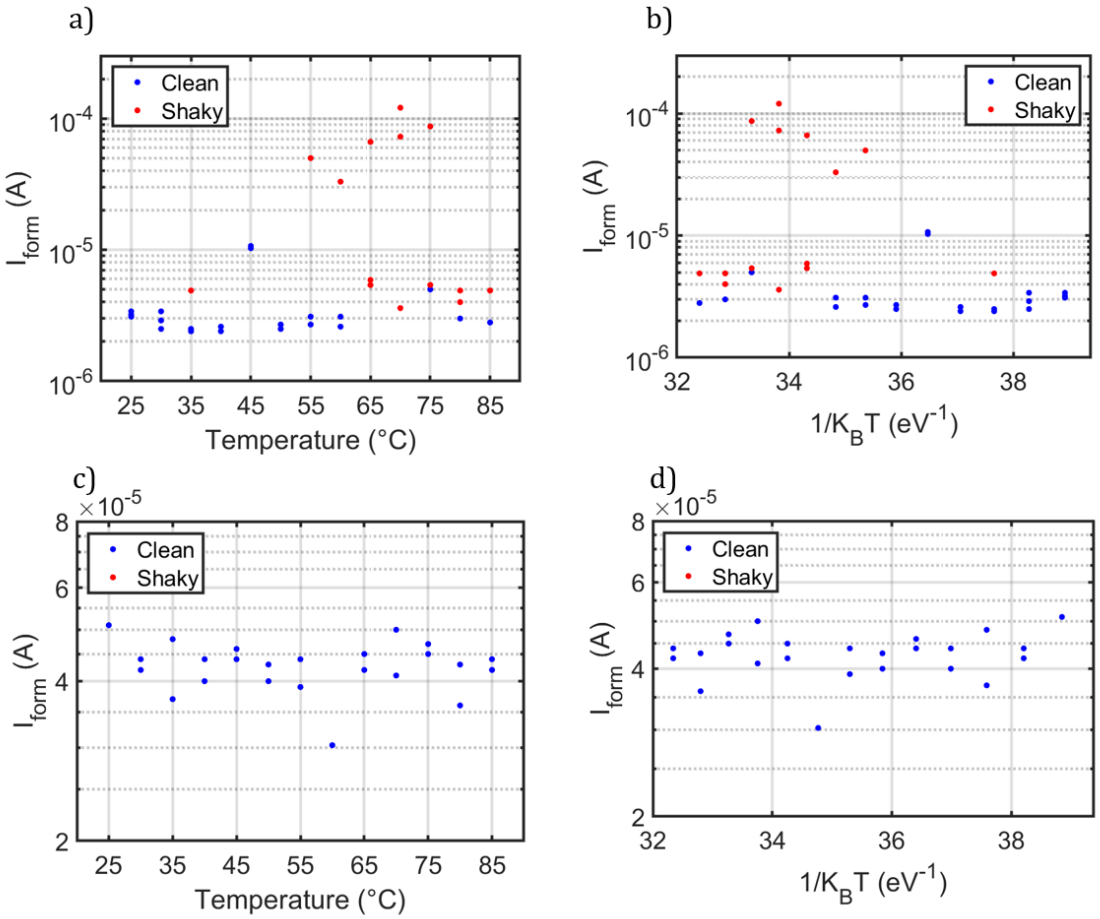

Figure 4.32: Forming current, I_{form}, as a function of the forming temperature and $1/k_BT$ for $HfO_2 - Ti$ devices of 12 µm size a),b), and 12 µm size c),d). The red dots highlight a "shaky", not in a single step, forming.

Activation energy extraction As described in Section 2.1, the forming process consists of the formation of oxygen vacancies - oxygen ions pairs from the HfO_2 layer. According to [31], the generation rate of stress-induced oxygen vacancy/ion pairs associated with the breakage of $Hf - O$ bonds can be described by Equation 4.1.

$$G_F(x,y,z) = \nu \cdot \exp\left(-\frac{E_A - b \cdot E(x,y,z)}{k_B \cdot T(x,y,z)}\right) \tag{4.1}$$

Where $\nu = 7 \times 10^{13}$ Hz is the effective vibration frequency of the $Hf - O$ bonds, $b = p_0 \cdot [(2+\epsilon_r)/3]$ the bond polarization factor, $p_0 = 5.2$ eA the molecular dipole moment, $\epsilon_r = 21$ the relative dielectric constant and E the electric field,[31].
Supposing that the generation rate G_F should be equal for devices on the same chip, it is possible to extract the activation energy E_A of the generation process by minimizing the standard deviation of $-\frac{E_A - b \cdot E(x,y,z)}{k_B \cdot T(x,y,z)}$. This is equivalent to minimizing std(G_F). The result of this calculation is shown in Figure 4.33, where it can be seen that the extracted activation energy $E_A \simeq 2.9$ eV which matches very well with the values in literature [31].

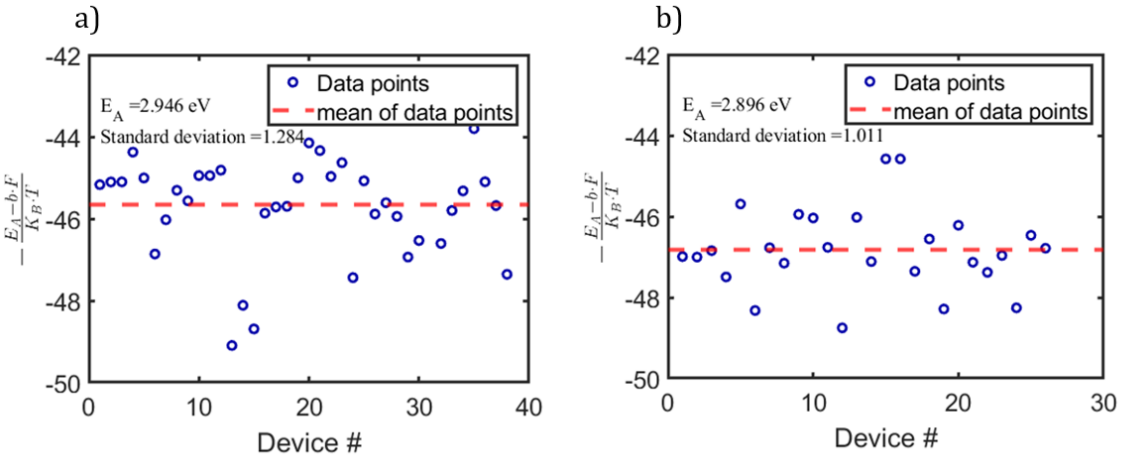

Figure 4.33: $-\frac{E_A - b \cdot E(x,y,z)}{k_B \cdot T(x,y,z)}$ plotted for all the measured devices of 12 µm size a), and 12 µm size b), calculated with the extracted value of the activation energy E_A.

4.3.4 Forming: HfO_2 – MOx technology

For HfO_2 – MOx devices the same trend of reduction of forming voltage with the increase of temperature is observed, but the slope is higher, indicating that for HfO_2 – MOx devices the forming voltage reduction with temperature is higher with respect to HfO_2 – Ti devices, Figure 4.34. The forming current, instead, has a clear trend: it increases with the temperature but not linearly. The forming current seems to saturate at both high and low temperatures, Figure 4.35.

Considering the MOx layer as a semiconductor, its resistance will decrease with increasing temperature, therefore more current can flow. The resistance reduction of the MOx layer results in a smaller voltage drop over the MOx layer and consequently an increase in the voltage drop over HfO_2, facilitating the forming process. These effects can explain the higer reduction of forming voltage with temperature and the trend of increasing forming current with temperature.

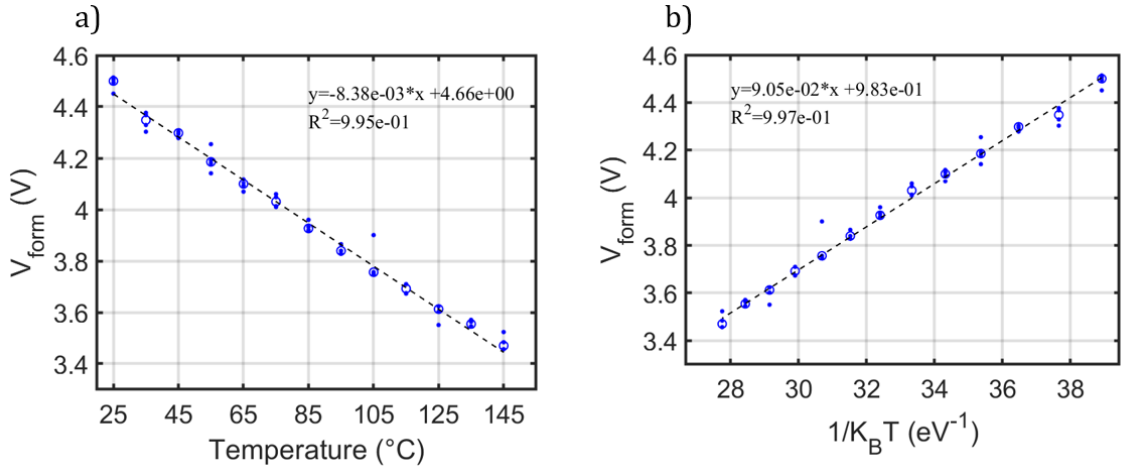

Figure 4.34: Forming voltage, V_{form}, as a function of the forming temperature and $1/k_BT$ for HfO_2 – MOx devices of 12 µm size.

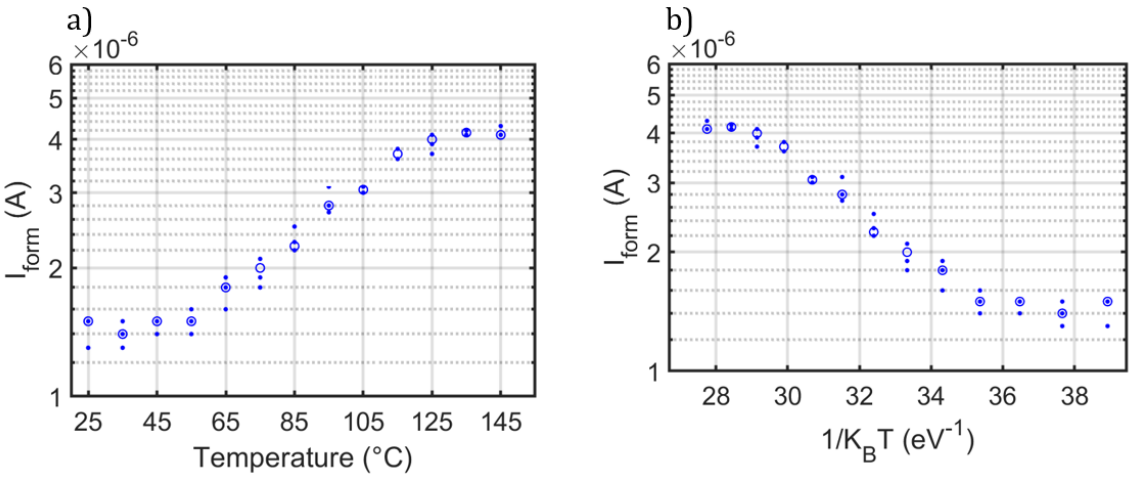

Figure 4.35: Forming current, I_{form}, as a function of the forming temperature and $1/k_BT$ for $HfO_2 - MOx$ devices of 12 µm size.

4.3.5 Cycling at room temperature after temperature dependent forming

Since high temperature forming brings a forming voltage reduction, it can be exploited to reduce the maximum voltage required for the operation of the devices, considering that the forming voltage is the largest voltage used for device operation. In active crossbars the memristor is in series with a transistor to facilitate the programming operations. In this context, relaxing the constraint on the maximum voltage enables the use of smaller and more power-efficient transistors.
To understand if this approach is feasible the devices of Section 4.3.3 and 4.3.4 that have experienced temperature dependent forming are cycled at room temperature. Their performances in terms of HRS and LRS resistances and On-Off ratio are reported in Figure 4.36. It can be observed that there is no evidence of a degradation in the performance of the devices subjected to temperature-dependent forming, thus enabling the use of this approach to reduce the forming voltage.

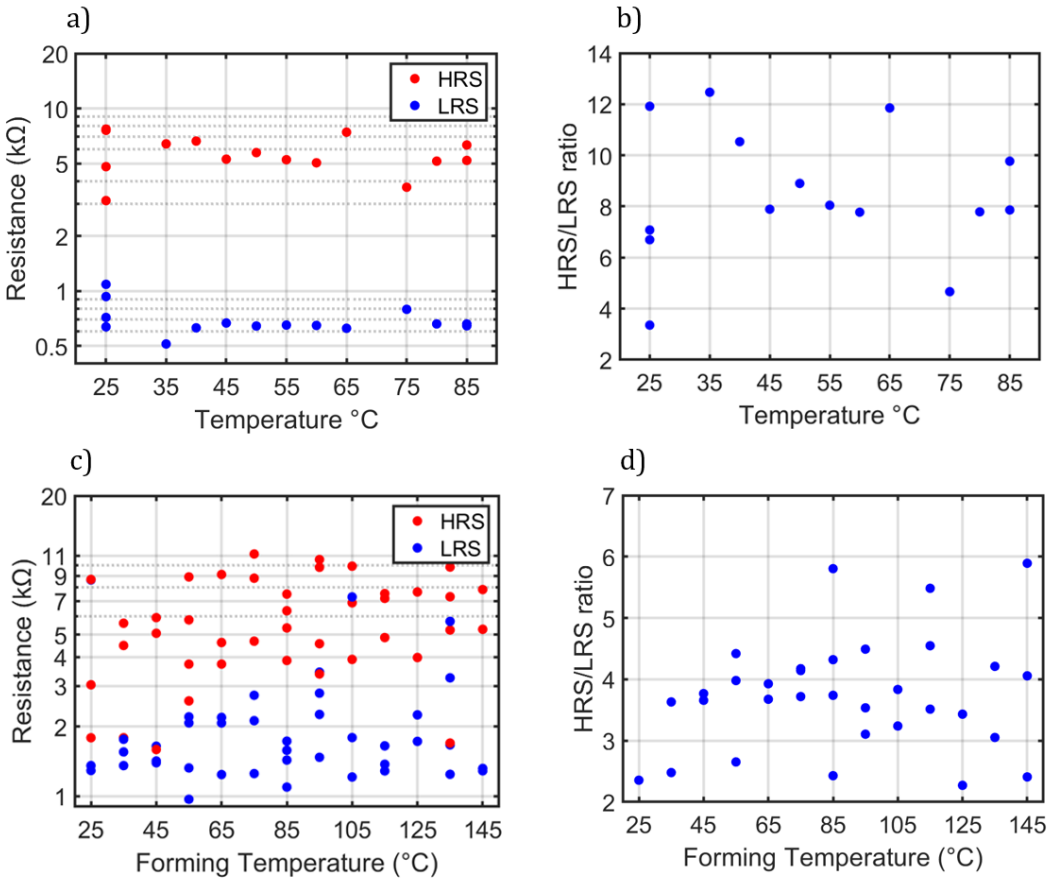

Figure 4.36: HRS and LRS resistances and On-Off ratio as a function of the forming temperature for $HfO_2 - Ti$ devices, a),b) and $HfO_2 - MOx$ devices, c),d).

4.3.6 "Hot forming"

In literature, [4, 34, 42],some articles report that a high temperature forming, "hot forming", which reduces the forming voltage, helps to improve some figures of merits of HfO_2-based RRAM devices such as the device-to-device variability, the stochasticity of HRS and LRS and the abruptness of the transitions, while others register no significant change, [32]. To quantify these possible improvements on the analyzed $HfO_2 - Ti$ technology, a dedicated experiment has been set up: the devices have been formed at 25 °C and at 150 °C and two device sizes have been measured: 60 µm and 30 µm. For each size and temperature 24 devices have been used.

From Figure 4.37 the only clear and consistent difference between the two forming temperatures is the higher LRS value for high forming temperature, which is consistent with a more controlled forming due to the smaller V_{form} and with [4]. However, the higher LRS value is gained at the expenses of higher device-to-device variability in the LRS value.
Figure 4.38 shows that the maximum change in resistance during the Set or Reset transition is not greatly affected by the "hot forming".
However, the standard deviation of the HRS, representing the cycle-to-cycle variability, shows a large reduction for high temperature forming, Figure 4.39, but at the same time the LRS shows a higher, but still small, variability. The higher variability in the LRS can be explained by its higher value represents a weaker filament in LRS, which is more susceptible to small changes.

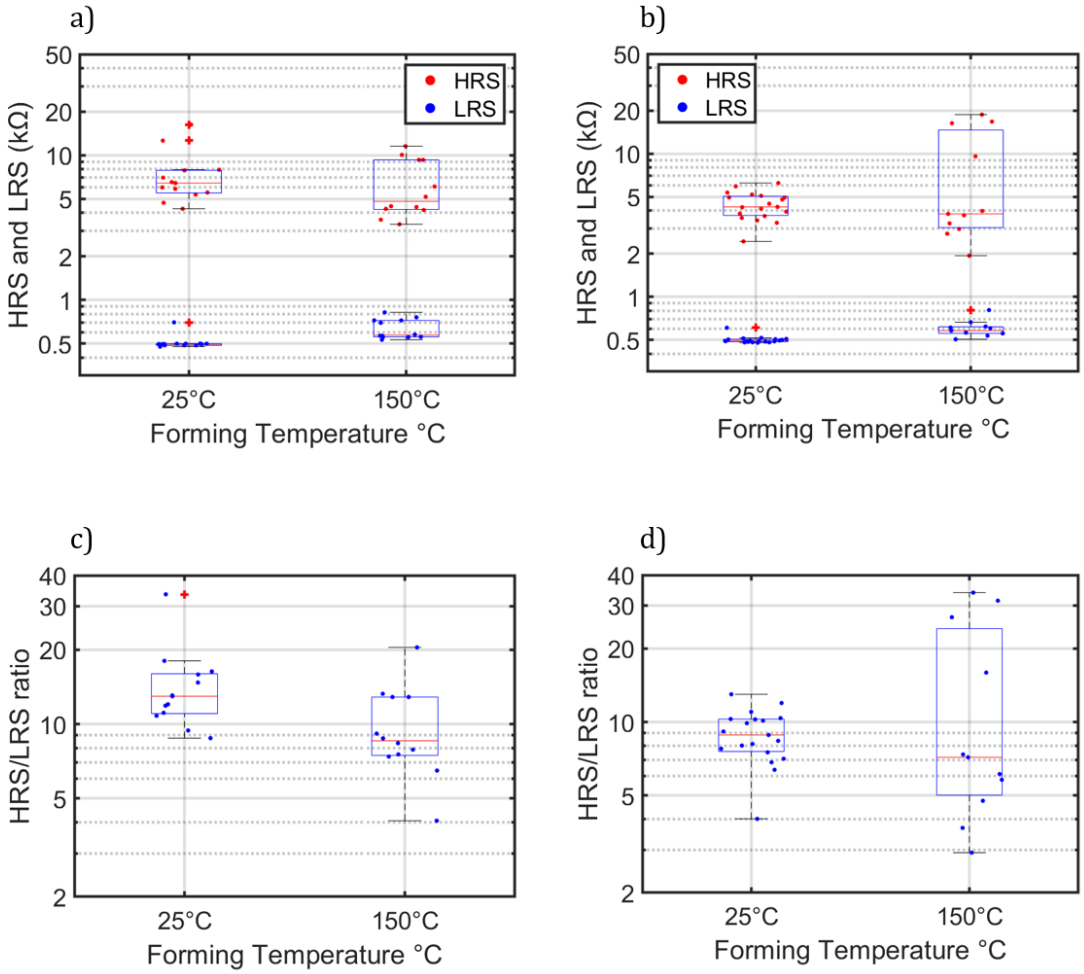

Figure 4.37: Boxplots representing the HRS and LRS resistances [a,b] and the On-Off ratio [c,d], on multiple devices grouped by forming temperature for [a,c] 60 µm devices, and [b,d] 30 µm devices.

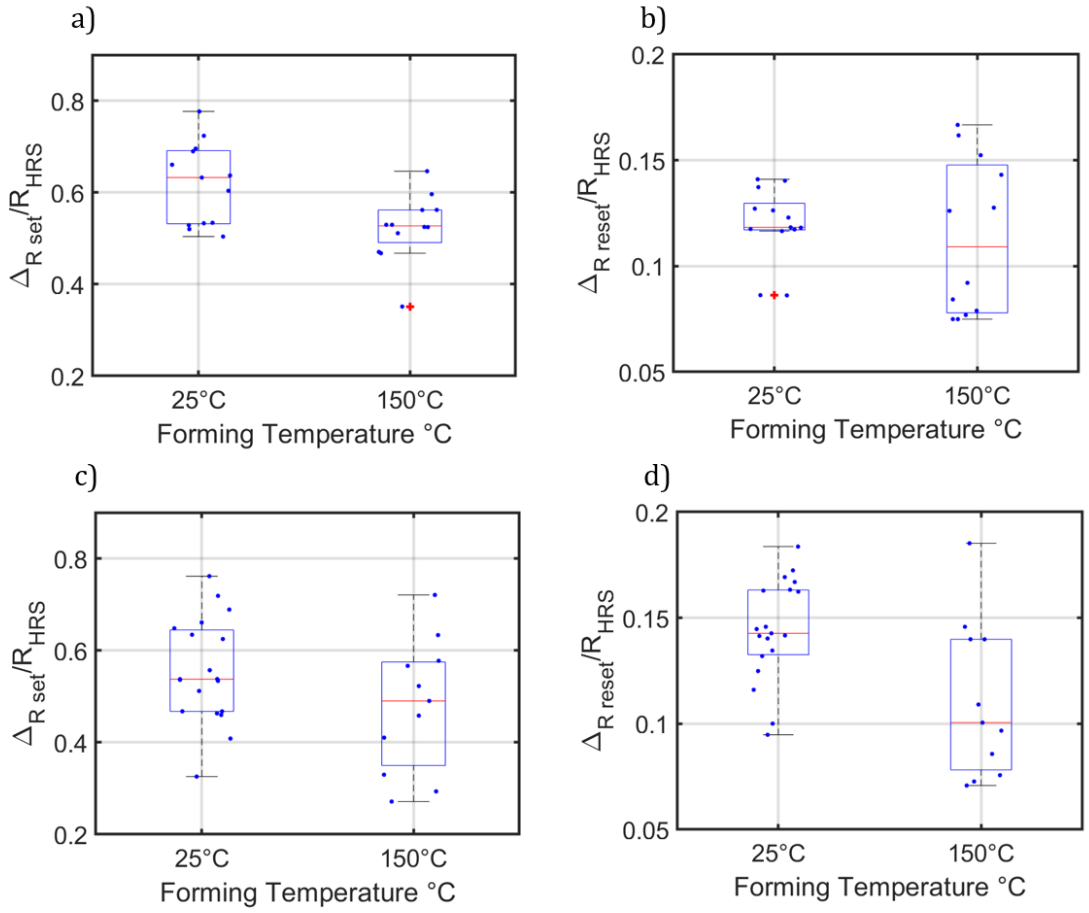

Figure 4.38: Boxplots representing the maximum change in resistance within 100 mV during the Set [a,c] or Reset [b,d] transition and normalizing it to the HRS value, on multiple devices grouped by forming temperature for [a,b] 60 µm devices, and [c,d] 30 µm devices.

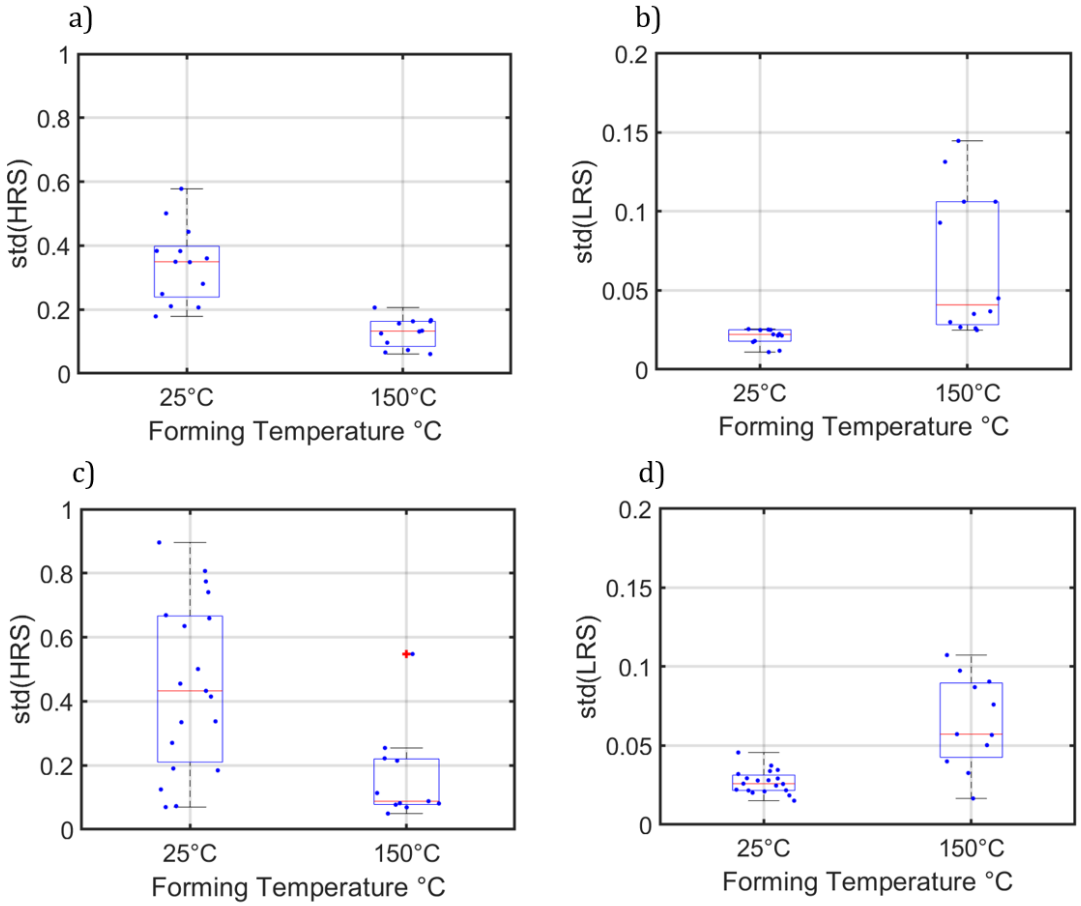

Figure 4.39: Boxplots representing the standard deviation of the HRS [a,c] and LRS [b,d] across the cycle performed, on multiple devices grouped by forming temperature for [a,b] the $HfO_2 - Ti$ technology, and [c,d] the $HfO_2 - MO_x$ technology.

4.3.7 HRS and LRS resistances

In this section the HRS and LRS resistances as a function of temperature are discussed, for devices which have been formed and cycled 10 times at room temperature before the start of the temperature-dependent experiment. In particular, for each technology, 5 devices in LRS and 5 devices in HRS have been measured. The resistance levels have been read 100 times at 0.2 V with an integration time of 0.1 s to obtain a very high accuracy. In Figure 4.40, examples of selected devices are shown.

As expected, since the devices are not metallic, the general trend is a reduction of the resistance with an increase in temperature. From Figure 4.41 it can be seen that while the LRS has a linear and consistent trend across all devices for both technologies, the HRS shows a consistent trend only in the $HfO_2 - MO_x$ technology. Moreover, the decrease of the slope of the LRS resistance is higher for $HfO_2 - MO_x$ devices with respect to $HfO_2 - Ti$ devices.
Baseline devices in LRS and $HfO_2 - MO_x$ devices in LRS or HRS have the conductive filament and the oxidized region on top of it in common. Thus, it is reasonable to assume that the consistent temperature-dependent behaviour is caused by one of those two elements, or both.

Another striking fact that it is clear in Figure 4.41 that the HRS resistance in $HfO_2 - MO_x$ devices drops extremely linearly and consistently across all the tested devices and it decreases much more than the LRS resistance.
In the proposed switching model, Section 4.1.3, in HRS $HfO_2 - MO_x$ devices have a larger, more oxidized region on top of the conductive filament, which is the major contribution to the HRS resistance. The more oxidized region can be the origin of the temperature dependent effect, since it is semiconductive, therefore its resistance is expected to drop with increasing temperature. This is consistent with the larger change of resistance of the HRS with respect to the LRS, because in LRS the more oxidized region is less extended and therefore contributes less to the total LRS resistance.

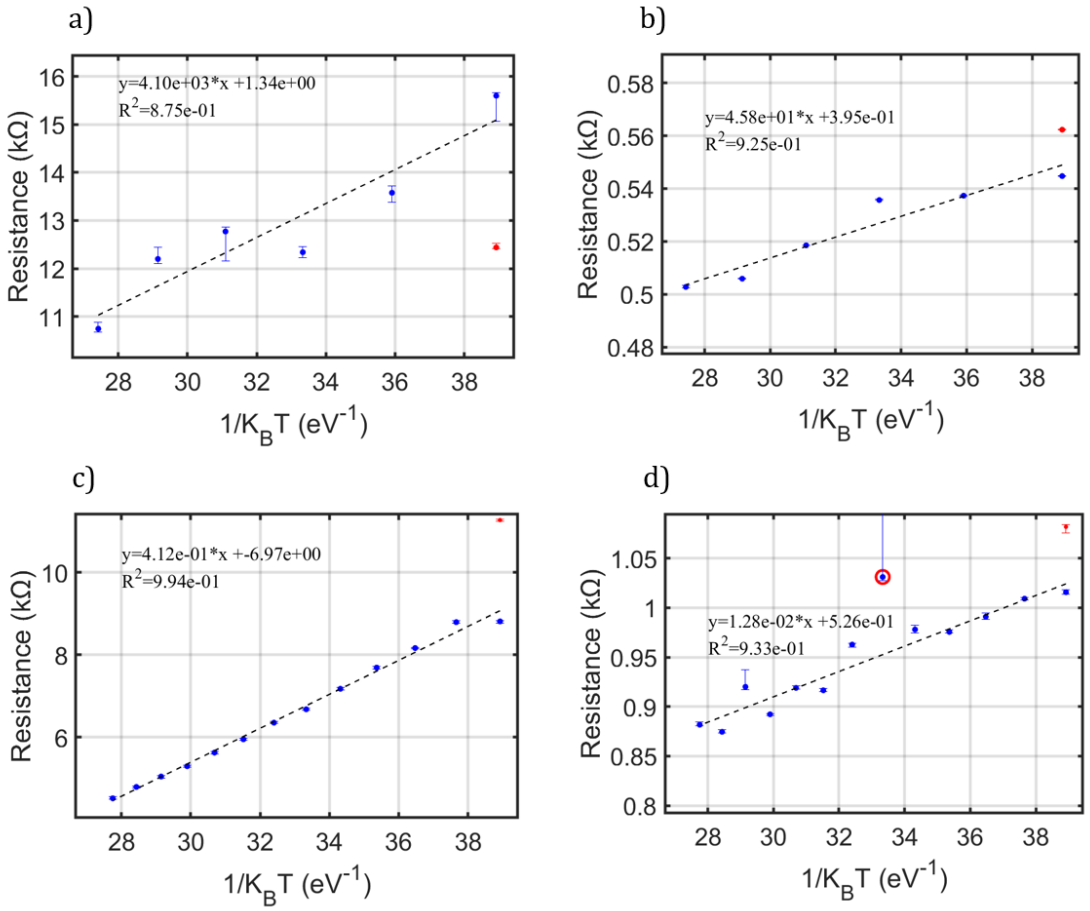

Figure 4.40: HRS [a,c] and LRS [b,d] resistances as a function of $1/k_BT$ for two examples of the [a,b] the $HfO_2 - Ti$ technology, and [c,d] the $HfO_2 - MO_x$ technology. The blue dots represent the average of the performed measurements, while the errorbars represent the maximum and the minimum value. The red dot represent the resistance measured at 25 °C after the temperature dependent experiment. The datapoints circled in red are caused by a bad contact due to a worn probe-tip. The linear fits are preformed on the average values excluding the outliers.

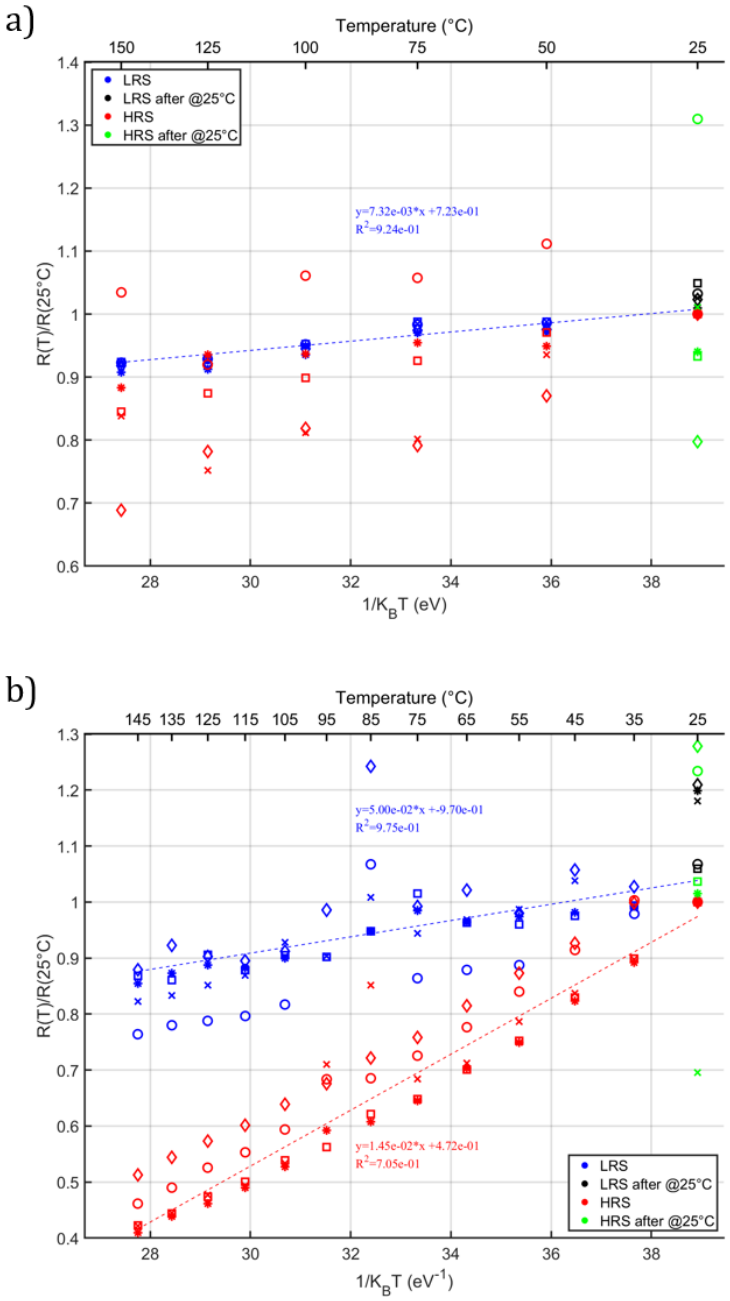

Figure 4.41: HRS and LRS resistances normalized to their value at 25 °C as a function of $1/k_BT$ for all measured devices of the a) the $HfO_2 - Ti$ technology, and b) the $HfO_2 - MO_x$ technology. Each shape represent a different device. In blue, the average LRS resistance at each temperature; in red, the average LRS resistance at each temperature; in black the average LRS resistance measured at 25 °C after the temperature dependent experiment; in green the average HRS resistance measured at 25 °C after the temperature dependent experiment.

4.3.8 Cycling

In this section the HRS and LRS resistance values of devices cycled 5 times at temperatures from 25 °C to 150 °C with steps of 25 °C are discussed. These devices have been formed and cycled 10 times at room temperature before the start of the temperature-dependent experiment. In particular, for each technology, 5 devices have been measured. In Figure 4.42 examples of selected devices are depicted.

For this experiment the devices are cycled and not just read at low voltages, thus a greater resistance change is expected with respect to the previous Section 4.3.7. At higher temperatures, the generation and the motion of the defects is facilitated, therefore the physical processes involved in the Set transition are enhanced and as a result the reduction of the LRS resistance is greater, Figure 4.43.

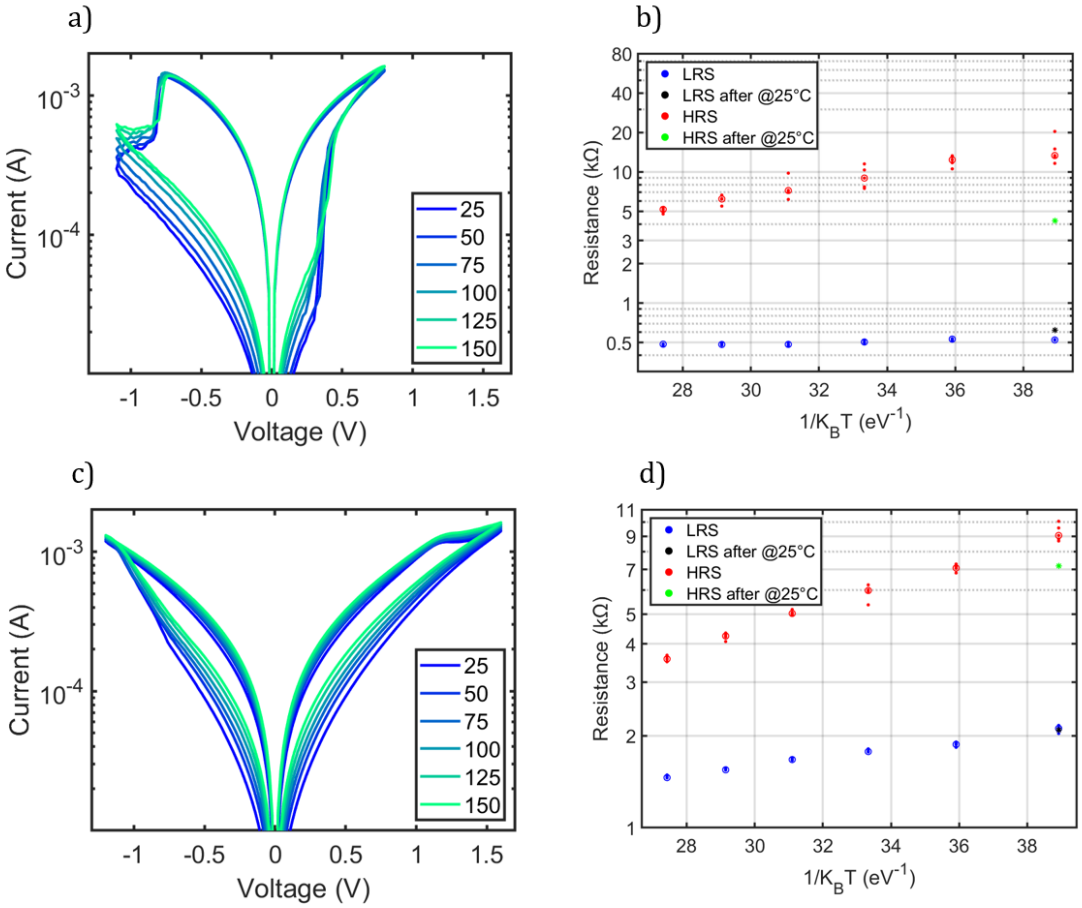

Figure 4.42: Cycling I-V curve [a,c] and recorded HRS and LRS [b,d] resistances during cycling as a function of $1/k_BT$ for one example of the [a,b] the $HfO_2 - Ti$ technology, and [c,d] the $HfO_2 - MO_x$ technology. In blue, the LRS resistance at each temperature; in red, the LRS resistance at each temperature; in black the LRS resistance measured at 25 °C after the temperature dependent experiment; in green the HRS resistance measured at 25 °C after the temperature dependent experiment.

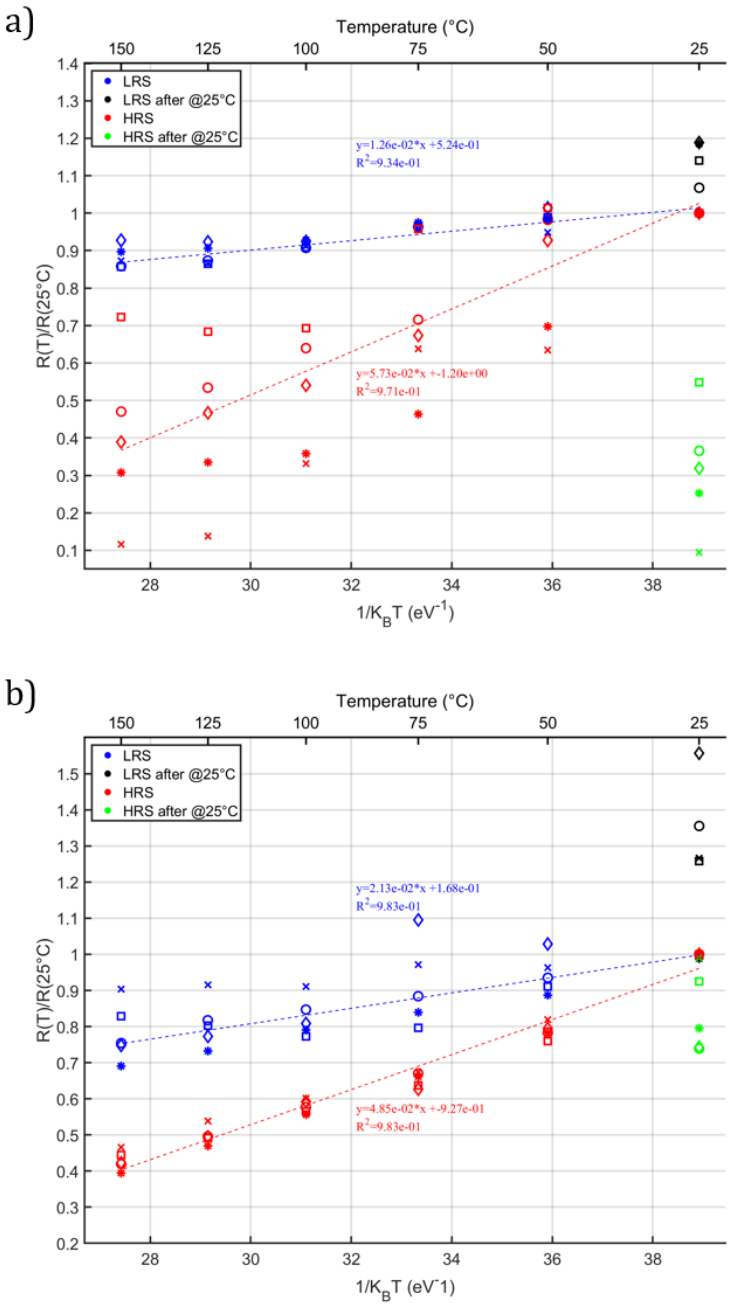

Figure 4.43: HRS and LRS resistances, recorded during cycling, normalized to their value at 25 °C as a function of $1/k_BT$ for all measured devices of the a) the $HfO_2 - Ti$ technology, and b) the $HfO_2 - MO_x$ technology. Each shape represent a different device. In blue, the average LRS resistance at each temperature; in red, the average LRS resistance at each temperature; in black the average LRS resistance measured at 25 °C after the temperature dependent experiment; in green the average HRS resistance measured at 25 °C after the temperature dependent experiment.

4.4 Ginestra

4.4.1 Modeling of HfO_2 – Ti devices

Using the Ginesta software it is possible to build a model of the $HfO_2 - Ti$ devices. To reduce the computation time it is necessary to approximate the actual device structure with a simplified model. In particular the device model has smaller dimensions, 50x50 nm^2, and it has two TiN ideal electrodes with in between 5.8 nm of HfO_2 and 2 nm of TiO_2. The small size of the simulated device can still accurately represent the actual device, since it has been shown that in pristine state the current scales perfectly with device area, Figure 4.20. Moreover, the active part of the device is in the HfO_2 layer and the small layer of oxidized Ti, which serves as oxygen reservoir.
The material parameters of the device model have been initially taken from the Applied Materials database, and fine tuned, using the parametric extraction tool, to match the experimental data in Figure 4.20.
The results are displayed in Figure 4.44, relevant material parameters used and extracted are: TiN work function $W - TiN = 4.5\,eV$, HfO_2 bandgap $E_g - HfO_2 = 5.8\,eV$, TiO_2 bandgap $E_g - TiO_2 = 1.2\,eV$, Oxygen vacancies in HfO_2 $d_{V+} - HfO_2 = 8.9 \times 10^{19}\,cm^{-3}$, mean of the thermal ionization energy of oxygen vacancies in HfO_2 $mean(E_{T,D(Q-1)} - HfO_2) = 1.89\,eV$, spread of the thermal ionization energy of oxygen vacancies in HfO_2 $spread(E_{T,D(Q-1)} - HfO_2) = 1.15\,eV$, phonon energy $E_{ph} - HfO_2 = 0.052\,eV$. These material parameters are coherent with the ones found in literature [31].

From Figure 4.44 it is clear that the simulation matches the experimental data well for the whole voltage range and for all temperatures. Moreover, this device model can also describe the subsequent operations performed on the device, such as forming and cycling, with the simulations outputs which are close to the actual experiment. This indicates that the physical description of the device, provided by the used material parameters and the simulation of the Ginestra software which includes all the conduction mechanisms in a systematic way, is very accurate.

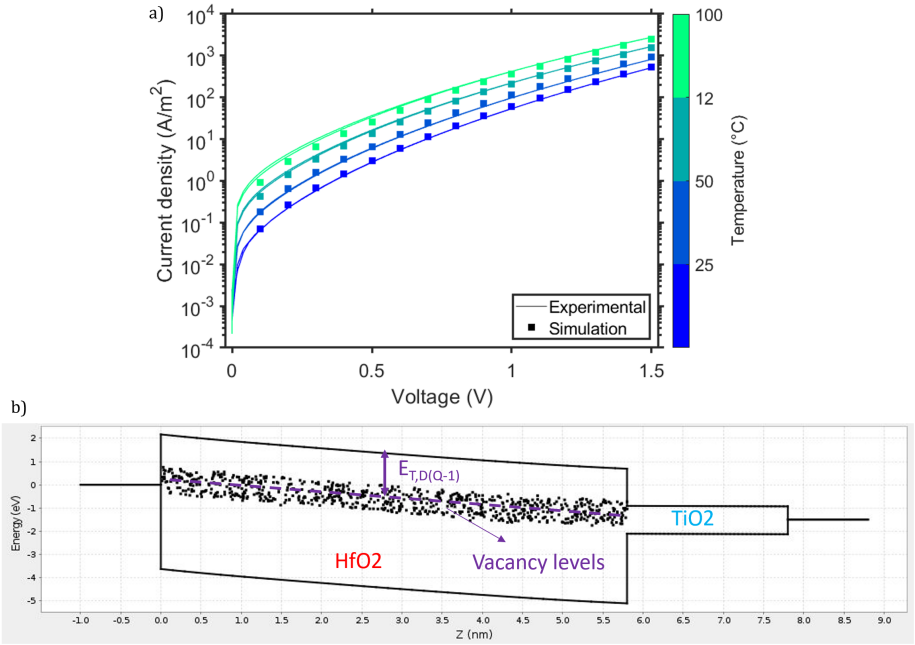

Figure 4.44: a) In squares the simulated J-V characteristic from the device model described in Section 4.4.1; in solid line the experimental J-V curve of $HfO_2 - Ti$. b) Band diagram of the device model described in Section 4.4.1 under positive bias. From left to right there is the ideal TiN bottom electrode, the HfO_2 layer, the TiO_2 layer and the ideal TiN top electrode.

4.4.2 Modeling of $HfO_2 - MO_x$ devices

Modeling of $HfO_2 - MO_x$ devices started with the device model of $HfO_2 - Ti$, because the HfO_2 layer is shared between the two technologies. The TiO_2 layer is taken as a template for the MO_x layer and it has been modified in its thickness, band gap, electron affinity and relative dielectric constant, with the values found in literature for the MO_x [38, 14, 11, 28].

MO_x model with oxygen vacancies defects Since MO_x is a sub-stoichiometric metal oxide, it has oxygen vacancies, therefore in the first modeling attempt oxygen vacancies have been added.

The dependencies of the pristine J-V curve on the defect density, energy position and spread, electron and hole effective mass, phonon energy and all the parameters adapted for the MO_x has been studied through multiple simulations.
Subsequently, using the parametric extraction tool those parameters have been fine-tuned to match as best the experimental data of 60 µm $HfO_2 - MO_x$ devices from ± 1 V to ± 3 V.
The 60 µm devices have been chosen because their pristine sweeps have higher currents, which are easier to measure. As shown in Figure 4.28, the leakage current does not scale perfectly with the device area, especially for low voltages. This finding suggests the presence of a preferential conduction path that is dominant at low voltages. However, with the Ginestra software it is not computationally possible to simulate devices as big as actual devices, thus the modeling and the simulations are limited to higher voltages where the conduction is mostly through the whole bulk and scales with area. For this reason the simulations on the $HfO_2 - MO_x$ devices start at ± 1 V.

In Figure 4.45 the results of the parametric extraction are shown. The extracted and used parameters are: MO_x bandgap $E_g - MO_x = 3.3\,eV$, MO_x electron affinity $E_A - MO_x = 3.1\,eV$, Oxygen vacancies in MO_x $d_{V+} - MO_x = 2.3 \times 10^{20}\,cm^{-3}$, mean of the thermal ionization energy of oxygen vacancies in MO_x $mean(E_{T,D(Q-1)} - MO_x) = 2\,eV$, spread of the thermal ionization energy of oxygen vacancies in MO_x $spread(E_{T,D(Q-1)} - MO_x) = 0.9\,eV$, phonon energy $E_{ph} - MO_x = 0.44\,eV$.
Figure 4.45 a clearly shows that this model is not able to fully describe the experimental data for positive and negative voltages. The model succeeds in matching the experimental data at positive biases while it completely fails at negative voltages.
From Figure 4.45 b and c it is possible to understand why the simulated J-V curve is so asymmetric. At positive voltages, Figure 4.45 b, the major contribution to the conduction is from the defects in the HfO_2 layer, follwoed by tunneling from those defects to the Conduction band of MO_x. This behaviour is highlighted by the fact that the current flowing through the defects in MO_x is very low and there is a higher current through the defects in the HfO_2, which are closer to the Conduction band of MO_x.
At negative voltages instead, Figure 4.45 c, the conduction is through the defects of both HfO_2 and MO_x.

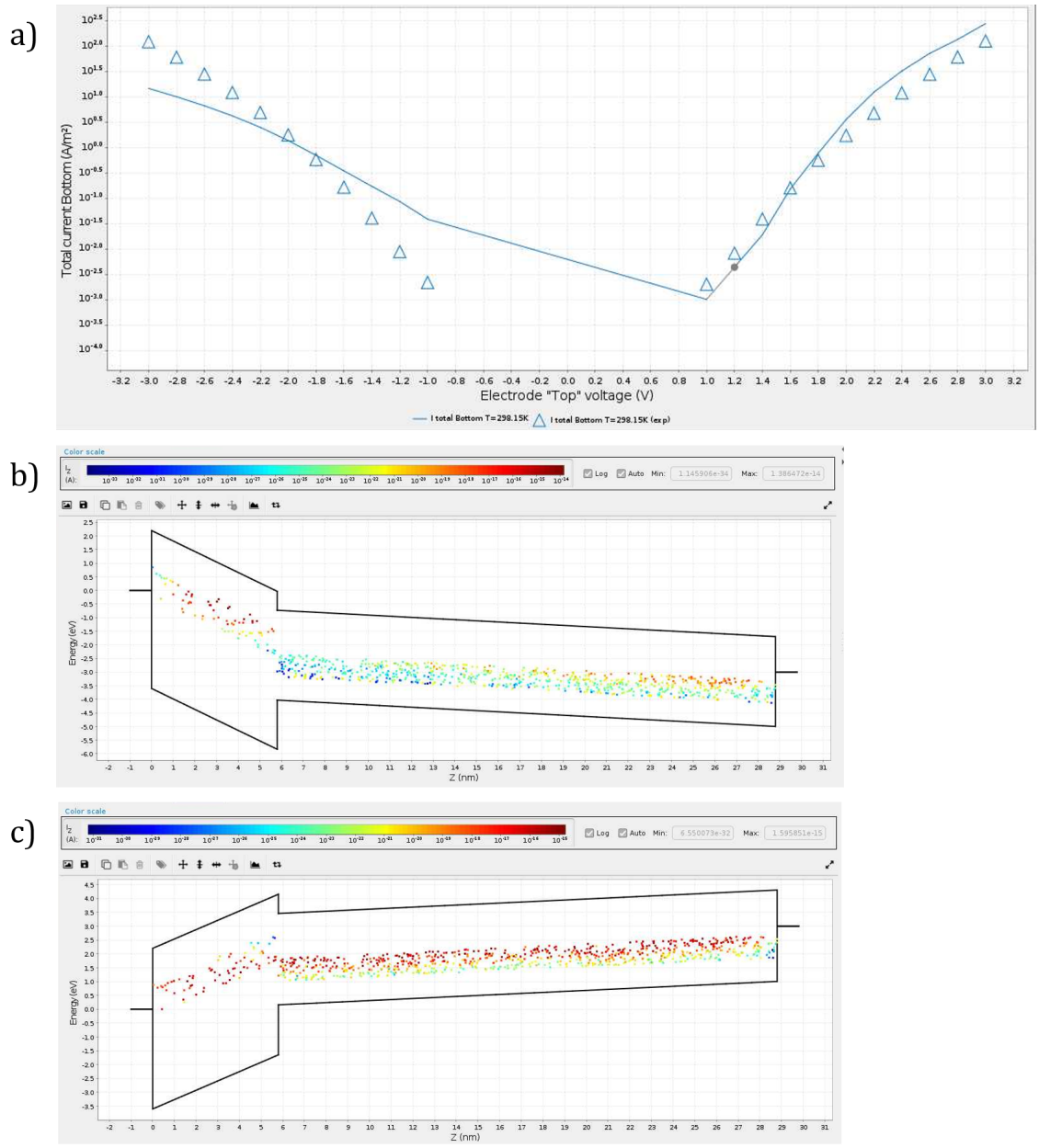

Figure 4.45: a) In solid line the simulated J-V characteristic from the device model described in Section 4.4.2; in triangles the experimental J-V curve of $HfO_2 - Ti$ sampled every 200 mV. b) Band diagram of the device model described in Section 4.4.2 at 3 V. c) Band diagram of the device model described in Section 4.4.2 at −3 V. The color mapping of the defects in b) and c) represent the current in flowing in the defects in the Z direction: trough the stack; red represent highest current, blue lowest current. From left to right there is the ideal TiN bottom electrode, the HfO_2 layer, the MO_x layer and the ideal TiN top electrode.

MO_x model without oxygen vacancies defects Since in the previous model the positive part of the J-V curve matched the experimental data and the negative experimental part was not radically different from the positive part, it is necessary to make the device model more symmetric. In particular, at positive biases the conduction is not defect assisted in the MO_x layer but is trough tunneling from the defects of the HfO_2 to the Conduction band of MO_x. Therefore, the new device model starts again from the $HfO_2 - Ti$ model with the modified TiO_2 layer, but without added defects.
The focus of the simulations has been on having a more symmetric conduction for positive and negative biases by favoring the tunneling from the defects in HfO_2 to the Conduction or Valence band of MO_x. This has been achieved by appropriately tuning the electron affinity and the bandgap in order to have a similar interface of the MO_x with the defects of the HfO_2.
Afterwards, parameters such as the electron and hole effective mass and tunneling effective mass, the electron and hole mobility and the relative dielectric constant have been modified to understand their effects on the conduction. Those tests highlighted the fact that at negative voltages there is mostly hole conduction in MO_x and at positive voltage conduction is mostly due to electrons.
As a final step an automatic parametric extraction has been used to fin- tune the relevant parameters in order to match at best the experimental data. For the same reasons explained in the previous paragraph, the experimental data from 60 µm $HfO_2 - MO_x$ devices from ± 1 V to ± 3 V has been used.

In Figure 4.46 the results of the parametric extraction are shown. The extracted and used parameters are: MO_x bandgap $E_g - MO_x = 2.6\,eV$, MO_x electron affinity $E_A - MO_x = 2.9\,eV$, electron tunneling effective mass $m^*_{e,T} - MO_x = 0.3$, hole tunneling effective mass $m^*_{h,T} - MO_x = 0.0014$.
The simulation in Figure 4.46 roughly matches the experimental data for the studied voltage interval and the three temperatures, However, it is necessary to specify that this is an incomplete model in that it only describes the part of the conduction that scales with device area.
From Figure 4.46 b and c it can be observed that the conduction is actually due to tunneling from the defects in the Conduction or Valence band, depending on the polarity of the pplied voltage.

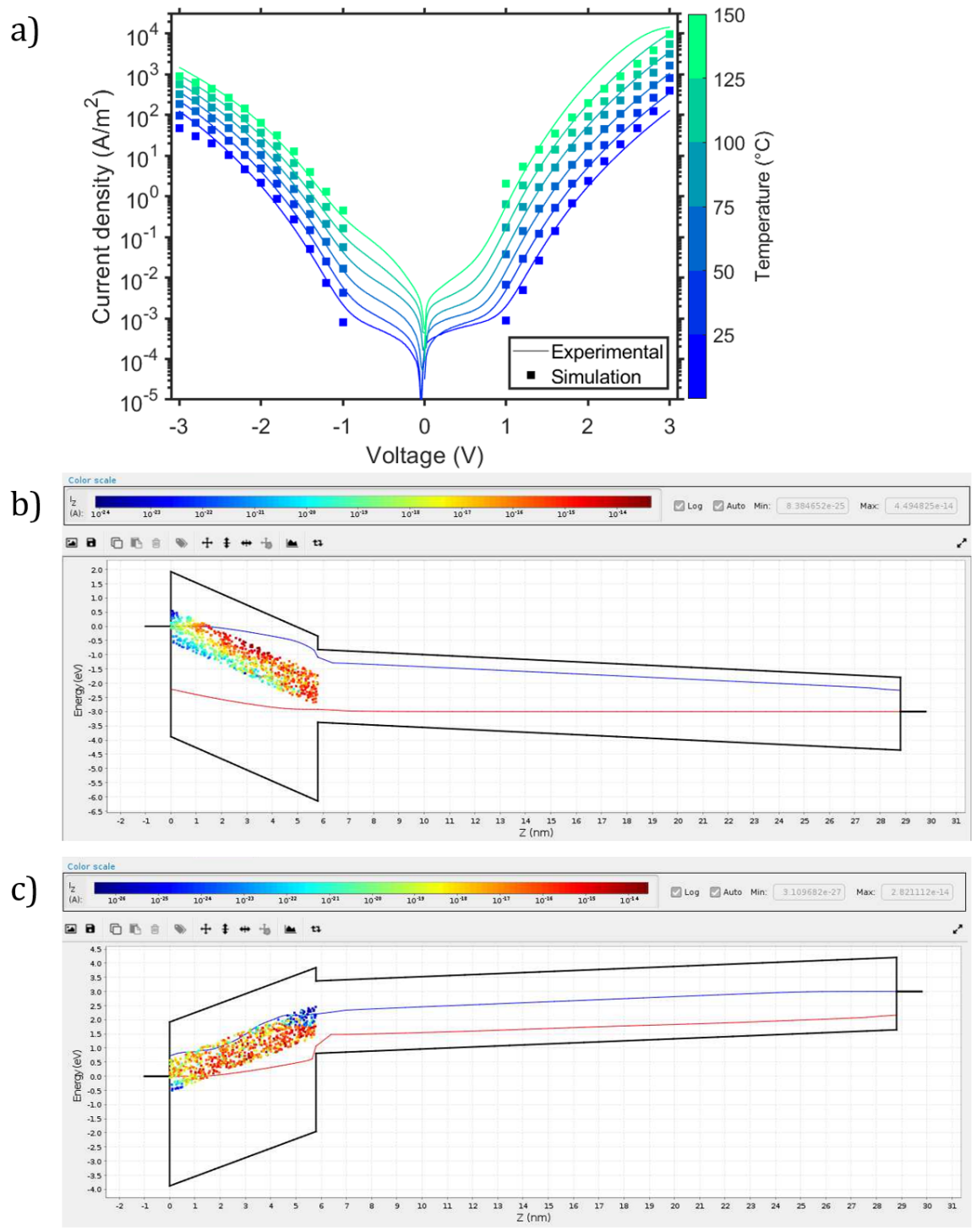

Figure 4.46: a) In squares the simulated J-V characteristic from the device model described in Section 4.4.2; in solid line the experimental J-V curve of $HfO_2 - Ti$. b) Band diagram of the device model described in Section 4.4.2 at 3 V. c) Band diagram of the device model described in Section 4.4.2 at -3 V. The color mapping of the defects in b) and c) represent the current in flowing in the defects in the Z direction: trough the stack; red represent highest current, blue lowest current. The blue line is the electron quasi Fermi level and the red line is the hole quasi Fermi level. From left to right there is the ideal TiN bottom electrode, the HfO_2 layer, the MO_x layer and the ideal TiN top electrode.

5 Conclusions

In this thesis a novel technology for the fabrication of RRAM devices, consisting of a $HfO_2 - MO_x$ material stack, has been analyzed and compared to a well-known RRAM technology found in literature, the $HfO_2 - Ti$ material stack. To goal was to assess the performance and to gain more understanding of the physical principles of operation. For the $HfO_2 - MO_x$ technology a switching model has been proposed that is centered around a more oxidized region that forms on top of the conductive filament. The proposed model can explain the opposite polarity of the Set/Reset transitions of the $HfO_2 - MO_x$ devices and the experimental data gathered throughout the experiments.

The performance was evaluated by means of DC and AC measurements on numerous devices of multiple sizes. It has been shown that while having a lower memory window, the $HfO_2 - MO_x$ technology can provide smoother transitions and has lower cycle-to-cycle variability. In the pulsed AC characterization, $HfO_2 - MO_x$ devices gradually changed their resistance while maintaining a good On/Off ratio. Moreover, the devices were able to reach a stable symmetry point, which can be used to enhance the training accuracy with the Tiki-Taka algorithm. However, the resistance of both technologies is too low for computational implementations in crossbar arrays, because the required power is too high and it would be difficult to accurately read such low resistances due to the additional resistances of the interconnections in the crossbar array. Thus, future research should focus on increasing the resistance values of the devices. Moreover, the endurance and the stability of the devices upon the application of a large number of pulses should be carefully investigated and improved if necessary.

The temperature-dependent experiments showed that in pristine state $HfO_2 - MO_x$ devices present a reduced leakage current with respect to the well-known technology, a hysteresis activated by positive voltages and asymmetry for positive and negative voltage biases. Moreover, the leakage current does not scale with the area at low voltages, suggesting that the conduction is through a preferential path that originates in the MO_x layer. These characteristics are absent in $HfO_2 - Ti$ devices: the devices are symmetric, they do not show hysteresis and as expected the leakage current scales perfectly with device area. These findings indicate that the MO_x layer has an active effect also in pristine state.
It has been shown that by high temperature forming it is possible to reduce the the forming voltage without altering the device performance during cycling.

By measuring the temperature dependence of the HRS and LRS it has been found that whether the device is cycled or not, the LRS decreases slightly with temperature for both technologies. Instead, the HRS shows a clear resistance reduction trend only for $HfO_2 - MO_x$ devices, suggesting a physical difference in the HRS between the two technologies. The large reduction of the HRS value with temperature is well-explained by the proposed switching model of $HfO_2 - MO_x$, since the more oxidized region on top of the filament can be considered semiconductive and therefore its resistance should decrease significantly with temperature.

Using the Ginestra software, the analyzed $HfO_2 - Ti$ devices have been modeled. It has been shown that the model can simulate and match the experimental data in pristine state for all the studied temperatures. Moreover, the same model is able to describe the subsequent operations of $HfO_2 - Ti$ devices, such as forming and cycling.
The $HfO_2 - Ti$ model has been used as a starting point for the simulations of the $HfO_2 - MO_x$, since the HfO_2 layer is shared. It is not computationally possible to simulate devices as large as actual devices, therefore the modeling and simulations of the $HfO_2 - MO_x$ devices are for voltages higher, in magnitude, than $\pm 1\,V$, where the conduction is mostly through the bulk and thus scales with area. With those considerations, a model was found that was able to roughly match the experimental data starting from $\pm 1\,V$. Yet, the model should be improved and expanded in order to explain also the hysteresis in pristine sweeps and the conduction through a preferential path that is dominant at low voltages.

References

[1] Applied mdlx™ ginestra™ simulation software.

[2] A. Berman. Time-zero dielectric reliability test by a ramp method, 1981.

[3] T. Bertaud, M. Sowinska, D. Walczyk, S. Thiess, A. Gloskovskii, C. Walczyk, and T. Schroeder. In-operando and non-destructive analysis of the resistive switching in the ti/hfo2/tin-based system by hard x-ray photoelectron spectroscopy. *Applied Physics Letters*, 101(14):143501, 2012.

[4] B. Butcher, G. Bersuker, K. Young-Fisher, D. Gilmer, A. Kalantarian, Y. Nishi, R. Geer, P. Kirsch, and R. Jammy. Hot forming to improve memory window and uniformity of low-power hfox-based rrams. In *2012 4th IEEE International Memory Workshop*, pages 1–4. IEEE, 2012.

[5] U. Celano, J. Op de Beeck, S. Clima, M. Luebben, P. M. Koenraad, L. Goux, I. Valov, and W. Vandervorst. Direct probing of the dielectric scavenging-layer interface in oxide filamentary-based valence change memory. *ACS applied materials & interfaces*, 9(12):10820–10824, 2017.

[6] T. Chang. *Tungsten Oxide Memristive Devices for Neuromorphic Applications.* PhD thesis, 2012.

[7] F.-C. Chiu. A review on conduction mechanisms in dielectric films. *Advances in Materials Science and Engineering*, 2014, 2014.

[8] F.-C. Chiu, H.-W. Chou, and J. Y.-m. Lee. Electrical conduction mechanisms of metal/ la 2 o 3/ si structure. *Journal of Applied Physics*, 97(10):103503, 2005.

[9] F.-C. Chiu, C.-Y. Lee, and T.-M. Pan. Current conduction mechanisms in pr 2 o 3/oxynitride laminated gate dielectrics. *Journal of Applied Physics*, 105(7):074103, 2009.

[10] L. Chua. Memristor-the missing circuit element. *IEEE Transactions on circuit theory*, 18(5):507–519, 1971.

[11] L. G. Gerling, S. Mahato, A. Morales-Vilches, G. Masmitja, P. Ortega, C. Voz, R. Alcubilla, and J. Puigdollers. Transition metal oxides as hole-selective contacts in silicon heterojunctions solar cells. *Solar Energy Materials and Solar Cells*, 145:109–115, 2016.

[12] T. Gokmen and W. Haensch. Algorithm for training neural networks on resistive device arrays. *Frontiers in neuroscience*, 14:103, 2020.

[13] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning.* The MIT Press, 2016.

[14] M. T. Greiner and Z.-H. Lu. Thin-film metal oxides in organic semiconductor devices: their electronic structures, work functions and interfaces. *NPG Asia Materials*, 5(7):e55–e55, 2013.

[15] C. Hamann, H. Burghardt, and T. Frauenheim. *Electrical conduction mechanisms in solids.* Vch Pub, 1988.

[16] M. Horowitz. 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 10–14. IEEE, 2014.

[17] D. Ielmini and R. Waser. *Edited by Daniele Ielmini and Rainer Waser Resistive Switching.* 2016.

[18] D. Ielmini and H.-S. P. Wong. In-memory computing with resistive switching devices. *Nature Electronics*, 1(6):333–343, 2018.

[19] D. S. Jeong, H. B. Park, and C. S. Hwang. Reasons for obtaining an optical dielectric constant from the poole–frenkel conduction behavior of atomic-layer-deposited hfo 2 films. *Applied Physics Letters*, 86(7):072903, 2005.

[20] G. K. Johnsen. An introduction to the memristor-a valuable circuit element in bioelectricity

and bioimpedance. *Journal of Electrical Bioimpedance*, 3(1):20–28, 2012.

[21] H. Kim, M. Rasch, T. Gokmen, T. Ando, H. Miyazoe, J.-J. Kim, J. Rozen, and S. Kim. Zero-shifting technique for deep neural network training on resistive cross-point arrays. *arXiv preprint arXiv:1907.10228*, 2019.

[22] D. Lamb. Electrical conduction mechanisms in thin insulating films, methuen and co. *Ltd., London*, page 56, 1967.

[23] L. Larcher. Statistical simulation of leakage currents in mos and flash memory devices with a new multiphonon trap-assisted tunneling model. *IEEE Transactions on Electron Devices*, 50(5):1246–1253, 2003.

[24] M. L. Lee, E. A. Fitzgerald, M. T. Bulsara, M. T. Currie, and A. Lochtefeld. Strained si, sige, and ge channels for high-mobility metal-oxide-semiconductor field-effect transistors. *Journal of Applied Physics*, 97(1):011101, 2005.

[25] W.-C. Lee and C. Hu. Modeling cmos tunneling currents through ultrathin gate oxide due to conduction-and valence-band electron and hole tunneling. *IEEE Transactions on Electron Devices*, 48(7):1366–1373, 2001.

[26] J. W. McPherson. Time dependent dielectric breakdown physics - models revisited. *Microelectronics Reliability*, 52:1753–1760, 9 2012.

[27] A. Mehonic, D. Joksas, W. H. Ng, M. Buckwell, and A. J. Kenyon. Simulation of inference accuracy using realistic rram devices. *Frontiers in neuroscience*, 13:593, 2019.

[28] M. Mews, L. Korte, and B. Rech. Oxygen vacancies in tungsten oxide and their influence on tungsten oxide/silicon heterojunction solar cells. *Solar Energy Materials and Solar Cells*, 158:77–83, 2016.

[29] N. F. Mott and E. A. Davis. *Electronic processes in non-crystalline materials*. Oxford university press, 2012.

[30] A. Padovani, L. Larcher, G. Bersuker, and P. Pavan. Charge transport and degradation in hfo 2 and hfo x dielectrics. *IEEE electron device letters*, 34(5):680–682, 2013.

[31] A. Padovani, L. Larcher, O. Pirrotta, L. Vandelli, and G. Bersuker. Microscopic modeling of hfo x rram operations: From forming to switching. *IEEE Transactions on electron devices*, 62(6):1998–2006, 2015.

[32] E. Perez, L. Bondesan, A. Grossi, C. Zambelli, P. Olivo, and C. Wenger. Assessing the forming temperature role on amorphous and polycrystalline hfo2-based 4 kbit rram arrays performance. *Microelectronic Engineering*, 178:1–4, 2017.

[33] M. J. Rasch, D. Moreda, T. Gokmen, M. Le Gallo, F. Carta, C. Goldberg, K. El Maghraoui, A. Sebastian, and V. Narayanan. A flexible and fast pytorch toolkit for simulating training and inference on analog crossbar arrays. In *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pages 1–4. IEEE, 2021.

[34] C. Sung, S. Lim, H. Kim, T. Kim, K. Moon, J. Song, J.-J. Kim, and H. Hwang. Effect of conductance linearity and multi-level cell characteristics of taox-based synapse device on pattern recognition accuracy of neuromorphic system. *Nanotechnology*, 29(11):115203, 2018.

[35] L. Vandelli, A. Padovani, L. Larcher, R. Southwick, W. Knowlton, and G. Bersuker. A physical model of the temperature dependence of the current through SiO_2/HfO2 stacks. *IEEE Transactions on Electron Devices*, 58(9):2878–2887, 2011.

[36] J. F. Verweij and J. H. Klootwijk. Dielectric breakdown i: A review of oxide breakdown, 1996.

[37] A. Wadsworth. *The Parametric Measurement Handbook 4th Edition*. Keysight, 2017.

[38] C. Walter, C. Hertzler, P. Devynck, G. Smith, and J. Peterson. Photodetachment of wo- 3: The electron affinity of wo3. *The Journal of chemical physics*, 95(2):824–827, 1991.

[39] G. Wentzel. Eine verallgemeinerung der quantenbedingungen für die zwecke der wellenmechanik. *Zeitschrift für Physik*, 38(6):518–529, 1926.

[40] E. Y. Wu and J. Suné. Power-law voltage acceleration: A key element for ultra-thin gate oxide reliability. *Microelectronics reliability*, 45(12):1809–1834, 2005.

[41] W. A. Wulf and S. A. McKee. Hitting the memory wall: Implications of the obvious. *ACM SIGARCH computer architecture news*, 23(1):20–24, 1995.

[42] X. Xu, L. Tai, T. Gong, J. Yin, P. Huang, J. Yu, Q. Luo, J. Liu, Z. Yu, X. Zhu, et al. 40× retention improvement by eliminating resistance relaxation with high temperature forming in 28 nm rram chip. In *2018 IEEE International Electron Devices Meeting (IEDM)*, pages 20–1. IEEE, 2018.

[43] F. Zahoor, T. Z. A. Zulkifli, and F. A. Khanday. Resistive random access memory (rram): an overview of materials, switching mechanism, performance, multilevel cell (mlc) storage, modeling, and applications. *Nanoscale research letters*, 15(1):1–26, 2020.

[44] M. A. Zidan, J. P. Strachan, and W. D. Lu. The future of electronics based on memristive systems. *Nature electronics*, 1(1):22–29, 2018.