Bipolar Resistive RAM based on HfO_2: Physics, Compact Modeling, and Variability control

Francesco Maria Puglisi, *Student Member IEEE*, Luca Larcher, *Member IEEE*, Andrea Padovani, *Member IEEE*, Paolo Pavan, *Senior Member IEEE*

***Abstract*—In this paper we thoroughly investigate the characteristics of the TiN/Ti/HfO_2/TiN Resistive Random Access Memory (RRAM) device. The physical mechanisms involved in the device operations are comprehensively explored from the atomistic standpoint. Self-consistent physics simulations based on a multi-scale approach are employed to achieve a complete understanding of the device physics. The latter includes different charge and ion transport phenomena, as well as structural modifications occurring during the device operations. The main sources of variability are also included by connecting the electrical response of the device to the atomistic material properties. The detailed understanding of the device physics allows developing a physics-based compact model describing the device switching in different operating conditions, including also the effects of cycling variability. Random Telegraph Noise (RTN), which constitutes an additional variability source, and its relations with cycling variability are analyzed. A statistical link between the programmed resistance and the worst-case RTN effect is found and exploited to include RTN effects in the compact model. Finally, we show how implementing an advanced programming scheme tailored on the device physics allows optimal control over variability and RTN, eventually achieving reliable and RTN-resilient two-bits/cell operations.**

***Index Terms*— RRAM, Resistive switching, high-κ dielectric, High Resistive State (HRS), Low Resistive State (LRS), Physics-based model, Compact model, Trap-Assisted Tunneling (TAT), Oxygen vacancy, Random Telegraph Noise (RTN).**

I. Introduction

The research of novel non-volatile memory technologies satisfying the stringent consumer market requirements has tremendously accelerated in the last years. Among the many proposed solutions, the Resistive Random Access Memory (RRAM) technology represents an attractive option due to its potential for low-complexity, high-density, high-speed, low-cost, low-energy non-volatile operation [1-9], which can be exploited in both embedded and stand-alone applications. Furthermore, RRAM devices can easily be integrated in the back-end of line (BEOL) thus enabling innovative logic-in-memory approaches and easier transition toward full-3D architectures [3, 9]. Within the large family of the metal-oxide-based resistance switching memories, a common characteristic is that their operating mechanisms involve the rearrangement of the dielectric material at the atomic level [10]. This is necessary to change the electrical resistance of the device between two distinct levels, i.e. the High- (HRS) and Low-Resistance-State (LRS), associated with different atomic arrangements of the dielectric material [4-10]. However, different materials for the switching layer and the electrodes result in different device behaviors [10], which are correctly described only by considering the material-specific characteristics and the interaction among the layers.

Here we consider transition metal-oxide (TMO) RRAM devices, specifically the bipolar TiN/Ti/HfO_2/TiN RRAM [6], which is very attractive since it can be easily integrated in the BEOL of the CMOS process. In addition, it offers excellent scaling perspectives – the material used for the switching layer (HfO_2) and for the electrodes (TiN) are already used in CMOS technology – it demonstrates excellent memory characteristics, data retention and endurance [9] with reduced implementation costs [1, 4, 6, 10]. Still, inherent stochastic features of this technology, i.e. randomness in the physical mechanisms involved in switching operations [11-15], are hampering its full industrial exploitation. Major concerns are represented by cycling variability and random telegraph noise (RTN) [11-15], which are hindering the device scaling and multi-bit storage implementation. So, the optimization of RRAM performance can be achieved only by understanding the physical processes responsible for variability and noise. To this point, it is mandatory to consider the HfO_2 material characteristics at the atomistic level. On the other hand, this approach requires achieving the self-consistent solution of a set of equations accounting for many physical processes occurring in the device [16], which results in a high computational burden. As such, it is compulsory to develop compact models able to capture the device characteristics, including variability and noise effects. These compact models have to be based on the device physics but must be slender enough to be used in circuit simulators, in order to enable variability-aware design of RRAM-based circuits [17-20].

In this paper, we analyze the physics of the device by providing a connection between its physical properties and the electrical measurements. We consider charge and ion transport and material-dependent structural modification occurring during RRAM operations. This allows describing critical RRAM features (i.e. switching characteristics, HRS and LRS

F. M. Puglisi and P. Pavan are with Dipartimento di Ingegneria "Enzo Ferrari", Università di Modena e Reggio Emilia, Via P. Vivarelli 10/1, 41125 Modena, Italy (e-mail: francescomaria.puglisi@unimore.it – phone: +39-059-2056324).

A. Padovani and L. Larcher are with Dipartimento di Scienze e Metodi dell'Ingegneria, Università di Modena e Reggio Emilia, Via Amendola 2, 42122 Reggio Emilia.

variability, noise) starting from its atomistic characteristics, i.e. the O-deficiency, O ion diffusion, and dielectric morphology (e.g. its grain structure). Then, we develop a compact model calibrated on the physics-based representation, providing a simple yet effective tool to be used in circuit simulators. The physical understanding of noise and variability allows including also these critical issues in the compact model, which can then be used to simulate the device behavior in RRAM-based circuits for innovative applications (e.g. logic, hardware security, neuromorphic circuits) [21-23]. The model is able to reproduce the experimental trends observed in RRAM devices with different geometry and dielectric stacks, operated under different voltage and temperature conditions. Finally, we show how an optimal programming scheme tailored on the device physics allows a refined control over noise and variability, enabling reliable multi-bit operation.

The paper is organized as follows. In Section II we review the state-of-the-art understanding of the physical mechanisms governing the RRAM operations and provide a comprehensive picture of the physical phenomena involved in the device operations. In Section III we derive and discuss the compact model which includes the critical variability issues. The RTN, its origin, and its relation with cycling variability are described in Section IV. In Section V we show how the implementation of a refined programming scheme enables reliable multi-bit operation with refined control over noise and cycling variability. Conclusions follow.

II. Physics-Based Simulation of HfO_2-RRAM

In recent years, significant research efforts resulted in a variety of RRAM models proposed in the literature, trying to capture the physics of these devices [24-28]. All the available approaches, though offering different physical perspectives, display some common features, which result from experimental evidences:

1. The device in pristine state is a leaky insulator, revealing area-dependent charge transport [24-28].
2. After performing a controlled breakdown, i.e. the forming process [24-28], charge transport shows no area dependence, due to the formation of a localized conduction path, i.e. the Conductive Filament (CF).
3. After the CF is formed, the device can be repeatedly switched between LRS and HRS through set and reset operations [24-28].

The forming operation is the preliminary step to enable resistive switching and is performed only once by applying a sufficiently high positive voltage at the top electrode while providing a current compliance to avoid the permanent device breakdown [24-28]. This results in the creation of a metallic-like CF [24-28] through the device structure, which shunts the two electrodes, provides a localized path for charge carriers flow, and determines the LRS. The reset operation, driving the device in HRS, is performed by applying at the top electrode a voltage pulse with opposite polarity compared to forming [24-28], and results in an atomic-level rearrangement of a portion of the CF, which changes its electrical properties. The set operation, driving the device back again in LRS, is performed by applying a pulse with the same polarity as the forming one (but with a smaller voltage) at the top electrode [24-28] and leads to the restoration of the after-forming CF characteristics [24-28]. For the sake of completeness, devices employing an ultra-thin HfO_2 layer (<3nm) may show forming-free characteristics [29-30], i.e. the device can be switched without performing the preliminary forming step. Although these common features, different approaches in the literature consider different physical mechanisms to be responsible for the set and reset operations. Roughly, existing models may be grouped based on their description of the physics of the reset operation, i.e. how a portion of the CF is rearranged at the atomic level, Fig. 1. Some models [24-26] associate the reset operation with the re-oxidation of a CF portion, Fig. 1 i,ii,iii), which creates a dielectric barrier within the CF, responsible for the resistance change. Conversely, other models [27] consider the CF as characterized by a constriction (its narrowest point), which implies quantization of the electrons wavefunction. In this framework, the resistance changes during set/reset transitions are described by modulations of the constriction width only, Fig 1 iv,v,vi). Also, some hybrid approaches trying to handle both scenarios are available in the literature [28]. Though both approaches may be used to reproduce the experimental trends, there is no ultimate experimental proof of the existence of a constriction in the CF. Similarly, there is no direct evidence of the formation of a dielectric barrier during the reset operation. As such, a clear picture of the device physics can be achieved only by employing self-consistent simulations of the atomistic processes occurring during forming, set, and reset operations.

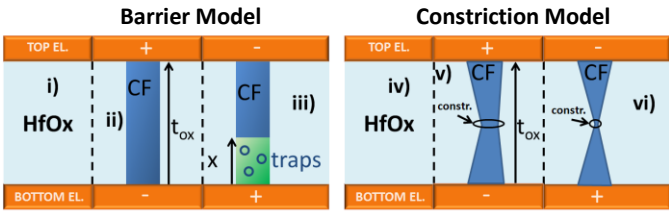

Figure 1. Schematic depiction of the constriction and the barrier models. i) The device in its pristine state. ii) After forming a CF is formed within the structure. Its shape is approximated as cylindrical. iii) The reset operation determines the creation of a dielectric barrier (green region) at the interface with the bottom electrode. iv) Pristine device representation (same as i). v) After forming, a filament with a constriction is formed within the dielectric layer. vi) The device resistance is modulated by the constriction width.

A. Simulation environment and device characterization

To achieve a thorough physical understanding of RRAM devices we perform measurements on devices from two different manufacturers. The devices from manufacturer "A" are TiN/5nm-Ti/5 or 7 nm-HfO_2/TiN with the Ti and HfO_2 layers created using Physical Vapor Deposition (PVD) and Atomic Layer Deposition (ALD), respectively. The devices from manufacturer "B" are TiN/10nm-Ti/10nm-HfO_2/TiN with the Ti and HfO_2 layers created using Sputtering and Atomic Vapor Deposition (AVD), respectively. The details of

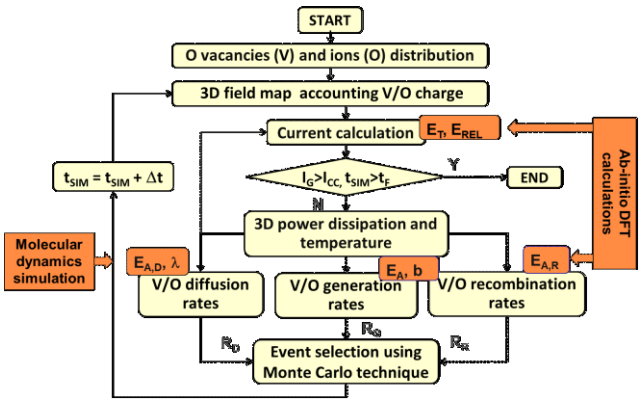

Figure 2. Flow chart illustrating the RRAM simulation system in MDLab. A multi-scale approach is used to describe the different mechanisms simultaneously occurring during device operations. The parameters defining the defects properties needed to calculate the TAT current and the generation, recombination, and diffusion rates are taken from ab-initio DFT and molecular dynamics simulations.

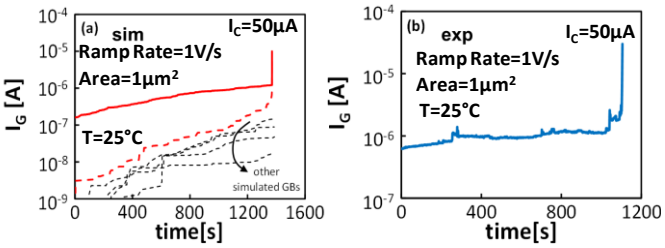

Figure 3. Simulated (a) and measured (b) evolution of the current during the forming operation on a 1μm^2 TiN/Ti/7nm-HfO$_2$/TiN RRAM device. A constant voltage of 2.45 V is applied. Simulations include the contributions of several GBs (dashed lines), but only one is eventually converted in the CF (dashed red line), which is responsible for the abrupt increase of the overall current (solid red line) at the end of the forming process.

the fabrication processes for both manufacturers are reported elsewhere [31-32]. Measurements were performed with standard lab equipment, specifically a Keithley 4200-SCS, used both in DC sweep and pulsed mode. Measurement results are then compared to simulations performed using MDLab software [33]. This tool includes several modules, Fig. 2, self-consistently describing the main physical mechanisms involved in RRAM stacks, as charge transport (including many conduction mechanism such as Trap-Assisted-Tunneling [34-35], and electrons drift across defect sub-bands [36]), generation, recombination, and diffusion of atomic species [37] (such as O ions and O vacancies). The software allows defining the physical properties of each layer composing the device, including defects, and simulates its electrical response considering the atomic properties of the materials through a multi-scale approach, Fig. 2. The software features full 3D description and a Monte-Carlo method to account also for the stochastic nature of some physical mechanisms involved in RRAM operations. Also, 3D temperature and potential maps of the device are available. In the following, we use the software to understand the physics of the RRAM operations. For consistency, we start the analysis by considering a device in its pristine state, Fig. 1 i,iv), performing the simulation of the forming process.

B. Forming operation

The forming process in HfO$_2$-based RRAM devices is typically described as the controlled breakdown of the oxide layer [10, 16-20, 24-28]. The description of this process requires considering the main charge transport mechanisms occurring in the dielectric layer, as well as the mechanisms responsible for structural modifications. From this standpoint, it has been assessed that charge transport in HfO$_2$ is mainly assisted by oxygen vacancy, Vo, defects [24, 34-35]. The physical mechanism involved may either be TAT [34-35] (when the Vo density is within a critical value) or an ohmic-like drift across a defect sub-band (when the Vo density exceeds a critical value) [24, 36]. As such, the response of the device in its pristine state depends on the initial Vo profile in the device, which requires careful consideration.

In crystalline HfO$_2$, or in grain bodies, the initial Vo profile results from imperfections in the manufacturing process. So, the Vo density is considered to be uniform and relatively low [38-41]. Amorphous films, conversely, may display a higher density of as-grown Vo. In the poly-crystalline HfO$_x$ layers (resulting from Post-Deposition Annealing of Atomic Layer Deposited HfO$_2$), the current flows preferentially at the grain boundaries, GBs, where the Vo tend to accumulate, as demonstrated by both experimental and theoretical studies [38-41]. Moreover, the Vo density at GBs is dramatically increased by the O extraction performed by the active metal layer (i.e. Ti in this case) overlaying the HfO$_2$ [42-44]. Experimental data indicate that inserting a thicker Ti above the HfO$_2$ layer allows extracting more oxygen ions from the oxide (probably form grain boundaries [24, 42-44]), generating more oxygen vacancy defects. This allows manipulating the device stoichiometry by simply adjusting the thickness of the Ti capping layer, thus modifying the forming voltage [6, 42-44]. Interestingly, it has been shown that the tuning of the Ti and the HfO$_2$ layer thicknesses also results in superior performance, e.g. larger reading window, extended endurance and reduced variability [42-44].

Experimental evidences show how the CF exhibits quasi-metallic I-V characteristics [6,7,10-12] indicating that it is formed by a localized oxygen-deficient region in the dielectric. The formation of the CF is thus believed to be the result of the metal-oxygen bonds breakage and the ensuing out-diffusion of the released O ions [24, 37]. Indeed, a bond breakage event results in the formation of an O ion and a Vo pair. Theoretical studies suggest that O ion diffusion in crystalline HfO$_2$ (or in a grain body) [38] is way more efficient than Vo diffusion [24]. Hence, the processes determining the structural modifications during the CF formation are described by considering charge transport together with both the bond breakage and the O ions diffusion (rather than Vo's). However, for completeness, calculations include also the possible Vo diffusion process.

The MDLab simulation environment calculates the Hf-O bond breakage probability at any location in the dielectric adopting the Mc-Pherson's thermochemical model [45]. The actual location of the breakage event is selected using the Monte Carlo method. The generation rate for Vo and O ion pairs, G, is described by:

$$G(x,y,z) = \upsilon e^{-\frac{E_A - b\cdot F(x,y,z)}{k_B \cdot T(x,y,z)}} \qquad (1)$$

The activation barrier, E_A, is lowered by the local electric field, F, induced by the polarization of the chemical bonds [16]. The strong dependency of the generation rate on both the local field and the temperature, T, is essential for the simulation of the forming dynamics, whose estimation requires calculating the temperature and the electric field evolution over time across the whole oxide. The temperature map is calculated starting from the power dissipated at each defect site (associated with the TAT charge transport) by solving the Fourier heat equation [16]. The 3D electric field profile is obtained by solving Poisson equation including ions and vacancies charges as well as the charge trapped at defect sites [16]. The software also includes the O ions and Vo diffusion and recombination. The parameters used to describe these phenomena are taken from ab-initio DFT and molecular dynamics simulations [38, 46-47], see Fig. 2.

The forming operation can be roughly considered as the sequence of two different phases, Figs. 3 and 4. In the former, the charge carriers flow mostly via pre-existing Vo (i.e. along the GBs in a poly-crystalline layer or within the whole structure for crystalline/amorphous oxides). The resulting current is assisted by TAT at Vo [34-35] (which can be regarded as isolated defects due to their relatively low density) and is generally low, as the associated power dissipation. As such, also the generation rate of new O/Vo pairs is low. Once a new O/Vo pair is generated inside the dielectric, the current increases since a new Vo is available to support the charge transport. The amount of increase depends on the energy and the position of the new Vo defect, according to the TAT model. In this phase the structural modifications arising from the defect creation are not spatially confined and the defect generation is a purely stochastic process.

However, when a critical number of Vo is generated close to each other, an Hf-rich cluster is created and the forming process enters the second phase, characterized by a field- and temperature-induced positive feedback. Here, the generation of Vo is self-accelerated [45], as a result of the local increase of field and temperature. The higher local current density associated with the Hf-rich cluster increases the temperature in the surrounding dielectric, favoring the generation of new defect pairs. This promotes the Vo generation around the initial cluster, the growth of which leads to the creation of extended defects and defect sub-bands [24]. Eventually, this mechanism results in the formation of a CF through the entire dielectric thickness. The overall picture is confirmed by simulations accounting for the contributions of several GBs in a poly-crystalline layer, Fig. 3. Indeed, even though in the initial phase the current flows through many GB paths, only the first GB entering the second phase will be converted in the final CF [16], Fig. 3. This means that the effect of the device geometry is negligible, as the active area of the device at which the switching and charge transport mechanisms occur

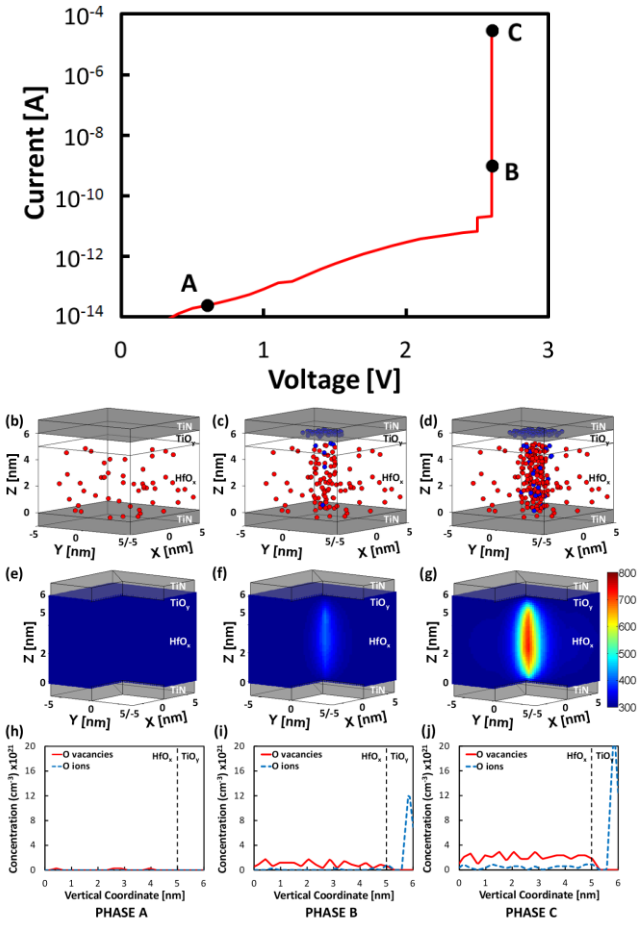

Figure 4. Simulation of the device evolution during ramped voltage forming (ramp rate=1V/s – T=25°C – I_C=10μA – 10x10nm^2 area) on a TiN/Ti/5nm-HfO_2/TiN RRAM device. The presence of the Ti layer results in the formation a thin TiO_y layer (included in simulations) during post-deposition annealing. a) The I-V characteristic. Three different snapshots are considered, namely Phase A, B, and C. b-c-d) The 3D distributions of oxygen vacancies (red, V^+) and ions (blue, O^{2-}) at the phases A, B, and C respectively. e-f-g) The associated 3D maps of the temperature induced by the power dissipation. h-i-j) The vertical profile of O vacancies and ions.

are extremely localized in the nano-sized CF, representing an excellent potential for the device scaling. In the framework proposed for the description of the CF formation, the grain boundaries serve as preferential initial points where O vacancy defects are either formed or accumulated. Nevertheless, this description can be applied also to amorphous materials, i.e. HfO_2, where, despite the more homogenous structure, microscopic differences in terms of atomic bond lengths and energies (e.g. observed using C-AFM) can originate preferential points/regions for charge transport and defect generation, serving as virtual GBs in the framework of the forming kinetics.

In Fig. 4 we report the detailed results of the simulation of the forming operation on a device with an amorphous oxide layer (i.e. the initial distribution of defect is uniform in the whole structure). Initially the power dissipation and the defect generation rate are low, phase A. However, when a defect cluster is randomly formed, phase B, the forming process enters the second phase which is characterized by a sudden localized temperature increase, culminating in the CF

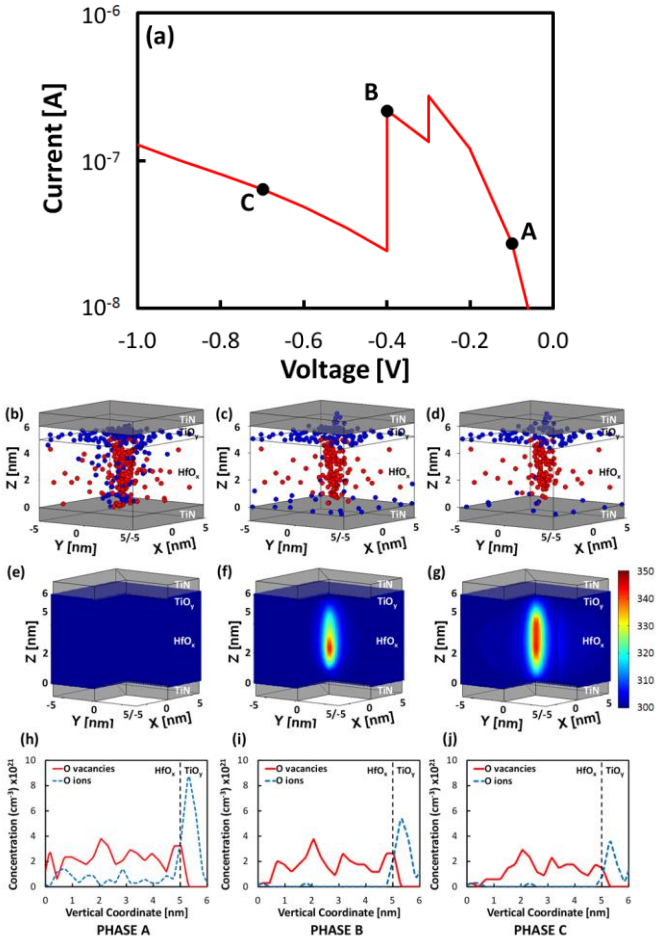

Figure 5. Simulation of the device evolution during the reset operation (ramp rate=1V/s – T=25°C – 10x10nm^2 area) on the device formed in Fig. 4. a) The I-V characteristic. Three different snapshots are considered, namely Phase A, B, and C. b-c-d) The 3D distributions of oxygen vacancies (blue, V^{2+}) and ions (red, O^{2-}) at the phases A, B, and C respectively. e-f-g) The associated 3D maps of the temperature induced by the power dissipation (much lower than in the forming operation). h-i-j) The vertical profile of O vacancies and ions. At phase C, the O vacancies density at the bottom interface is decreased, which reflects the formation of the dielectric barrier.

formation. The resistance of the resulting CF is controlled by the magnitude of the maximum current allowed to flow through it, i.e. the current compliance, which sets the CF cross-section [16]. Indeed, the dynamics of this positive feedback loop predicts that only imposing a current compliance limit allows interrupting the continuous O-depletion of the CF region eventually leading to destructive breakdown (no switching can happen).

The different phases of the forming process are associated with different dominant charge transport mechanisms. Initially, the TAT at Vo defects is responsible for the majority of the current flow. However, when a critical Vo density is reached the description of the current flow as assisted by TAT through isolated defects is no longer valid and the increasing delocalization of electrons among adjacent defects up to the formation of a sort of defect sub-band has to be considered. In this scenario, charge transport is described according to Landauer formalism [48], which is based on the quantum

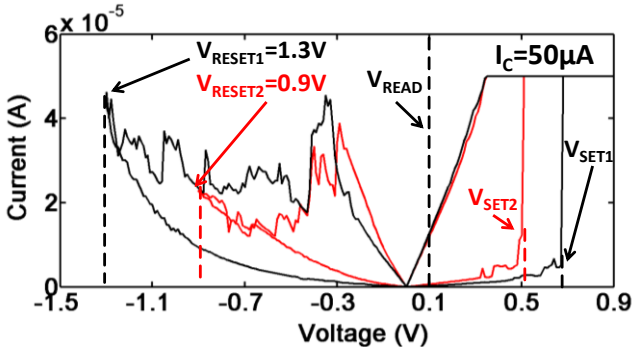

Figure 6. a) Typical I-V curves during switching at different V_{RESET}. The most important parameters controlling the device behavior are highlighted, together with the reading voltage, V_{READ}. Current is limited by compliance (I_C=50 µA). T=25°C. The ramp rate is 1V/s. The device has a TiN/5nm-Ti/5nm-HfO_2/TiN structure.

conductance. The O ions out-diffused during the forming operation accumulate at the top interface and in the surroundings of the CF, becoming available for the following switching operations (set/reset), see phase C in Fig. 4.

C. Reset and Set operation

Recognition of the regions where O ions accumulate during forming, set, and reset operations, which is crucial to model and understand the physical mechanisms ruling the reset operation, is still an open issue in the scientific community. The options are that O ions accumulate at either (i) the interface with the top electrode and/or (ii) at the CF surroundings, Fig. 4. The reset operation is performed by applying a negative voltage to the device (opposite to the voltage polarity applied during forming). This should drive the O ions from the top HfO_2 interface and from the surroundings of the CF back to the interface with the bottom electrode [24]. Eventually, the O ions should reach the interface with the bottom electrode, which is typically made of a material representing a diffusion barrier for O ions, like TiN. So, once the O ions reach the interface with the bottom electrode, they cannot further diffuse into it. Thus, the O ions accumulated at the bottom interface are randomly driven by both mutual Coulomb interaction and the density gradient to spread around and recombine with the Vo constituting the CF [24], Fig. 5. As such, the bottom tip of the CF starts being re-oxidized. The recombination between the Vo constituting the CF and the O ions continues as long as there are O ions available at the interface with the top electrode and/or in the CF surroundings to be driven by the applied voltage toward the bottom interface. This results in the formation of a dielectric barrier at the bottom of the CF, which is responsible for the resistance increase associated with the reset operation [24-26]. This is evidenced in Fig. 5, in which the I-V curve simulated upon the application of a voltage ramp is shown, together with schematic illustrations of the CF state (i.e. concentration of Vo and interstitial O ions) at three different stages of the reset operation. Noticeably, during a ramped voltage reset operation two concurrent and opposite effects occur. On the one hand, the voltage increase over time determines a higher current flow. On the other hand, the same voltage increase pushes more O ions toward the bottom electrode and promotes a more

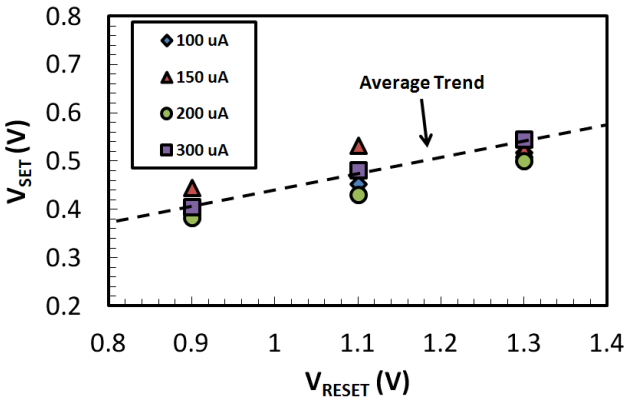

Figure 7. Experimentally retrieved set voltage vs. reset voltage characteristic shows a linear trend and no compliance dependence. Each point is the average on 5 TiN/5nm-Ti/5nm-HfO_2/TiN devices (different areas ranging from 100x100nm^2 to 1μm^2) and 9 switching cycles. T=25°C.

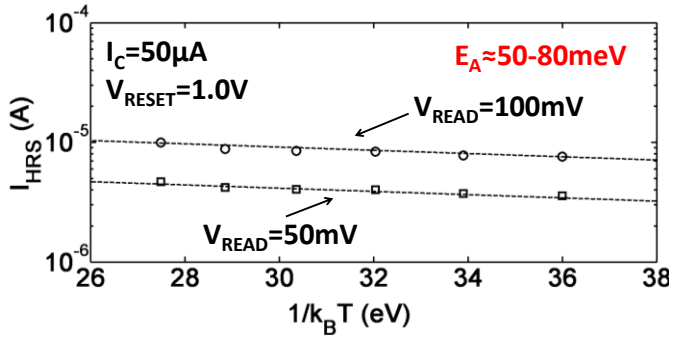

Figure 8. Arrhenius plot of the HRS current at two different V_{READ} (symbols) experimentally retrieved from a TiN/5nm-Ti/5nm-HfO_2/TiN device. The effective activation energy E_A≈50-80meV is extracted. All the symbols collected at the same reading voltage are related to the same bit, which is read at different temperatures. The device is a 100x100nm^2 cross-bar TiN/5nm-Ti/5nm-HfO_2/TiN formed at T=25°C. Ramp rate is 1V/s.

effective recombination (which is a partially stochastic process) with the Vo (filament re-oxidation), which results in the dielectric barrier formation and the consequent resistance increase. These two concurrent and competing processes result in the glitches observed between phases A and B in Fig 5.

Naturally, since the diffusion of the O ions and their recombination with Vo are partially stochastic processes, not all the Vo constituting the CF will recombine with the incoming O ions. As such, the reset dielectric barrier cannot be modeled as a pure HfO_2 layer, but as a Vo defect-rich insulating barrier [16]. These O-vacancy defects assist the TAT charge transport through the barrier, which represents a bottleneck for charge transport in HRS. If no defects were present in the barrier, the charge transport would be dominated by direct tunneling of the electrons across the barrier. Instead, in HRS, the charge transport in the barrier has been proven to be dominated by the electron TAT through Vo defects which did not recombine during the reset operation [16, 34].

The partially stochastic nature of the physical processes involved in the reset operation results in a significant cycle-to-cycle variability of the HRS resistance, which is due to the randomness of the recombination and diffusion processes [16, 24-26]. Indeed, after each reset cycle (performed in the same conditions) the effective barrier thickness may be different, as well as the position of the Vo defects within the barrier [16], which heavily influence the TAT charge transport through the barrier [34] resulting in a significant HRS resistance variability.

The set operation, typically performed by imposing the same current compliance as in the forming operation, allows the resistance switching from HRS to LRS, which corresponds to the restoration of the initial CF [24-26]. The set process is physically analogous to the forming one, discussed in Section II-B. It requires indeed breaking the barrier formed during reset, which happens at a much lower voltage, compared to the forming one. This is due to the high electric fields associated with the ultra-thin barrier. Thus, applying a low voltage results in a high electric field in the barrier, which determines a fast Hf-O bond breakage that restores the initial post-forming shape of the CF.

III. Compact Model

The complete atomistic description of the RRAM device shown in Section II can be used to support the experimental characterization of these devices. Nevertheless, this powerful representation is not best-suited for circuit simulations, which require more slender approaches. On the other hand, several reasonable simplifying assumptions can be made both in HRS and LRS to devise a physics-based compact model of the RRAM device. The model describes the switching operations including also cycling variability, which is a major concern for RRAM reliability and future developments. The proposed approach relies on the empirical representation of a key device parameter, i.e. the device resistance, linking the structural properties of both the CF and the barrier [49] to the electrical characteristics of the device [50-51]. The model correctly reproduces the behavior of the device, accounting for the effects of different current compliance, reset conditions, and temperature, as successfully verified against experimental data measured on RRAM devices with different geometries and produced by different manufacturers.

A. The Empirical Resistance Model

The empirical model describes the resistance in LRS and HRS in reading conditions and at room temperature (RT). It is based on the fact that the forming process can be described simply by the formation of one single CF shunting the two electrodes, which sets the LRS resistance [16, 24-26]. Due to the metallic-like characteristics of the CF, the LRS resistance can be modeled as:

$$R_{LRS} = \rho_{Hf,CF}\frac{t_{ox}}{S} \qquad (2)$$

where $\rho_{Hf,CF}$ is the resistivity of the CF, t_{ox} is the hafnium oxide layer thickness, and S is the cross section of the CF, which depends on the current compliance level [16, 24]. A larger current compliance during forming results in a larger filament cross section and a lower LRS resistance [24], assuming a constant value for the CF resistivity. The value of the CF resistivity has been estimated in the literature [52-53] by combining physical analysis (such as C-AFM scanning)

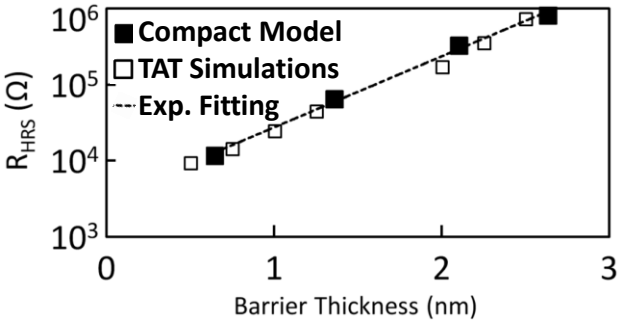

Figure 9. HRS resistance values vs. barrier thickness calculated at room temperature using the compact model (black full squares) and TAT simulation (black hollow squares), along with exponential fitting (dashed line). Here we consider a CF with a 10nm^2 cross-section and an oxide layer thickness of 5nm.

aimed at determining the CF cross-section and measurement of the CF resistance after forming [52]. In other works it is derived by matching noise data with theoretical models [53]. Typical values reported in the literature range from few kΩ·nm to some tens of kΩ·nm. Here we assume a value of 30 kΩ·nm. The LRS resistance temperature dependence is included in the resistivity, which weakly depends on the temperature, in agreement with the results reported in [24]. The HRS resistance depends instead both on the characteristics of the CF before the reset operation (i.e. its cross section) and on the reset voltage magnitude, V_{RESET}. Here V_{RESET} is defined as the maximum absolute value of the voltage applied during the reset operation, Fig. 6. As reported in the literature [51], a larger reset voltage implies the formation of a thicker barrier, due to the re-oxidation of a wider portion of the CF. This assumption is confirmed by two experimental evidences:

1) The LRS resistance remains almost unchanged (discarding very small variations) after set/reset operations (imposing the same current compliance, i.e. 50 μA in Fig. 6), confirming that set is related to the restoration of the same CF [24-26, 51].
2) The set voltage, V_{SET}, defined as the voltage at which the transition from HRS to LRS occurs during the set process, Fig. 6, linearly increases with the reset voltage, V_{RESET}, Fig. 7. This is fully consistent with the description of the set operation as a breakdown of the dielectric barrier—larger reset voltages require higher set voltages to reach the critical breakdown field across the barrier, i.e. to set the device [24]. Moreover, V_{SET} shows no dependence on the current compliance (or, equivalently, on the CF cross-section), Fig. 7, in agreement with the description of the set operation.

As such, the set voltage is expected to be proportional to the thickness of the barrier through the critical field (which is considered a constant depending on the material, neglecting interface effects): this suggests a linear relation between the barrier thickness and the reset voltage. Moreover, this is quantitatively consistent with experiments and physics-based simulations. Indeed, physics-based simulations show that the barrier resistance exponentially depends on its thickness while

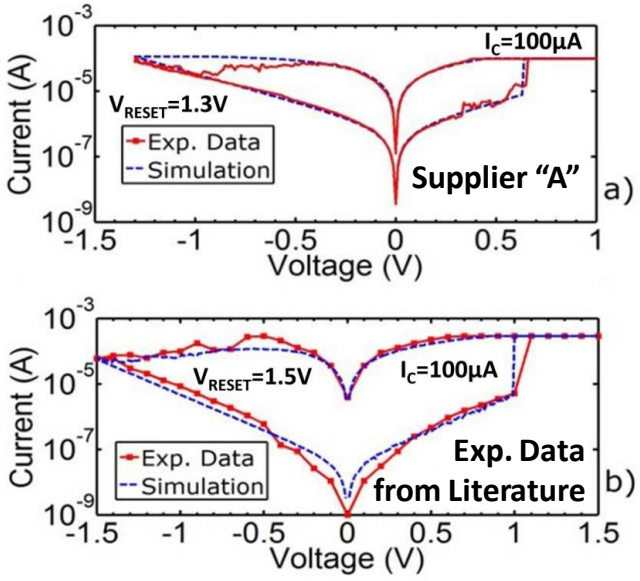

Figure 10. Experimental (red curve) and simulated (blue curve) I-V switching cycle curves for two different 1μm^2 TiN/5nm-Ti/5nm-HfO_2/TiN devices formed with I_C=100μA and reset with V_{RESET}=1.3 V (a - supplier "A") and V_{RESET}=1.5 V (b - experimental data from [50]). T=25°C and ramp rate=1V/s. Figure adapted from [51]. The simulation is performed always with the same model parameters values.

experimental results show an exponential increase of the device resistance with the absolute value of the applied voltage during the reset operation, see Fig. 9.

In HRS, the RRAM resistance can be described as the sum of two contributions: the former is given by the equivalent resistance of the barrier, having thickness x, and the latter is given by the resistance of the unbroken part of the CF, having thickness (t_{ox}–x), see also Fig. 1 ii,iii). The resistance of the unbroken part of the CF is given by (2), replacing t_{ox} by (t_{ox}–x). The resistance of the barrier exponentially depends on x through a normalization factor κ. The value (κ ≈ 0.42 nm) is consistent with TAT calculations at room temperature. The effect of temperature on the barrier resistance is included through an Arrhenius-like term modeling the temperature dependence of the TAT current, which is due to the lattice relaxation and the electron-phonon interactions [34-35]. The activation energy extracted from experimental data is E_A≈50-80 meV, Fig. 8, in agreement with the values reported in the literature. The complete model of the RRAM resistance, valid in both HRS and LRS, is thus given by:

$$R = R_{LRS}\left[\frac{t_{ox}-x}{t_{ox}} + \left(e^{\frac{x}{\kappa}}-1\right)e^{\frac{E_A}{k_B T}}\right] \qquad (3)$$

R_{LRS} can be simply extracted from I–V measurements in LRS. Both R and R_{LRS} are measured at the reading voltage of V_{READ}=100 mV and they are constant provided that the reading voltage does not exceed V_{READ}≈300 mV, which is usually avoided to prevent structural changes or undesired switching. Figure 9 shows the HRS resistance values calculated as a function of x using the compact model. Simulations performed using the physics-based TAT model in

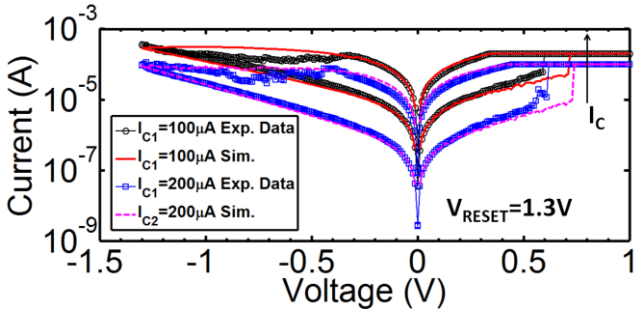

Figure 11. Experimental (solid curves with symbols) and simulated (lines w/o symbols) I-V switching cycle curves for two 1μm² TiN/5nm-Ti/5nm-HfO_2/TiN devices formed at two different compliances, I_{C1}=100μA (black circles and red solid line) and I_{C2}=200μA (blue circles and violet dotted line). T=25°C. Ramp rate=1V/s.

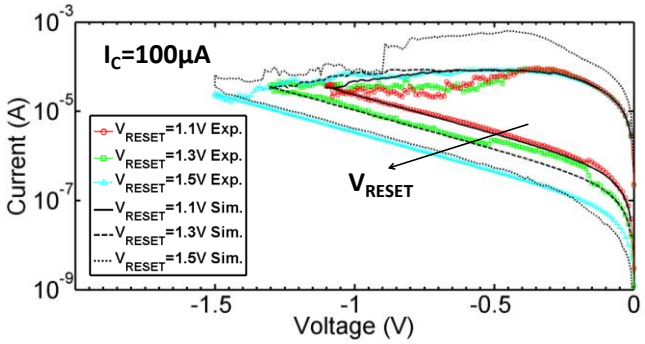

Figure 12. Experimental (solid curves with symbols) and simulated (black curves w/o symbols) I-V reset sweep curves for a 1μm² TiN/5nm-Ti/5nm-HfO_2/TiN device formed at I_C=100μA and reset at three different V_{RESET}. V_{RESET1}=1.1 V (red squares and solid line), V_{RESET2}=1.3 V (green squares and dashed line) and V_{RESET3}=1.5 V (cyan triangles and dotted line). T=25°C. Ramp rate=1V/s.

[34] (considering a trap density $N_T=2\cdot10^{21}$ cm^{-3}, in agreement with independent ab-initio calculations [49]) are also shown. The accurate matching among the results in Fig. 9 confirms the validity of the compact model, which can be used to link the electrical and physical characteristics of the device.

B. Resistive Switching

The empirical resistance model proves very useful to quickly characterize the physical parameters of the memory cell, starting from simple I-V measurements. In addition, the compact model also allows performing simulations of the dynamics of the switching operations. This allows reproducing the I-V curves for device manufactured with different process technologies, see Fig. 10, accounting for the effects of current compliance, reset conditions, temperature and cycle-to-cycle variability on the device. The development of a compact equations set describing the switching dynamics requires separately considering the set and the reset processes. To this point, it is useful to consider the device in LRS, right after forming. Indeed, the LRS resistance is mostly determined by the current compliance imposed during forming [51, 54]. This one-to-one correspondence seems to be valid in most of metal-oxide-based RRAMs, regardless the other forming conditions (e.g. temperature) [54]. As such we can safely discard the details of the forming operation in this compact representation. To model the reset process, we assume a linear relation between the barrier thickness and the absolute value of the applied voltage [55]. Reproducing the reset dynamics requires considering the evolution of the barrier thickness over time during the reset process, modeled as:

$$\begin{aligned} &if\ |V| > |V_{INIT}| \\ &\frac{dx}{dt} = R_{LRS}c_{xv}\cdot|V - V_{INIT}| \\ &else\ \frac{dx}{dt} = 0 \end{aligned} \qquad (4)$$

where c_{xv} is a coefficient and V_{INIT} the voltage at which the filament starts to be effectively oxidized. Interestingly, the linear dependence of the rate of the barrier thickness growth on the applied voltage is consistent with the effect of the voltage on the O ions, which oxidize the bottom tip of the CF. It is noteworthy that (4) may be refined to include second-order effects without changing the general arrangement of the model. In this representation, eq. (4) intrinsically includes the effect of the filament properties, e.g. cross-section, on the barrier growth dynamics through the R_{LRS} parameter. The value of the current I is calculated as a function of the barrier properties and the applied voltage through:

$$I(x,V) = I_0(x)\cdot\sinh\left(\frac{V}{V_0}\right) \qquad (5)$$

which models the non-trivial voltage dependence of the HRS current associated with the TAT charge transport. Interestingly, eq. (5) comprises two contributions: the former considers the effect of the barrier thickness on the current value, while the latter includes the voltage dependence [16, 55]. V is the applied voltage, V_0 is a fitting parameter which can be extracted from the experimental data. I_0 depends on the barrier thickness, which changes during reset, and is calculated as:

$$I_0(x) = \frac{V_0}{R(x)} \qquad (6)$$

where R(x) is the resistance of the device in reading conditions (i.e. at read voltage $V_{READ}\approx0.1$ V), calculated through the empirical resistance model in Section III-A, which includes thermal effects. The barrier growth process is considered to end when the negative voltage ramp reaches -V_{RESET}: when the absolute value of the voltage reduces below V_{RESET}, both the barrier thickness and I_0 remain unchanged, reflecting the fact that the barrier is not altered to any further extent. This is confirmed by the I-V characteristics in HRS, which are indeed accurately described through (5) assuming a constant barrier thickness and I_0.

The set operation, similar to the forming one, is modeled as a field-induced breakdown of the barrier created during reset [16, 24-26]. The barrier breakdown can be modeled in a compact manner by considering the value of the electric field within the barrier. When the electric field overcomes a critical value, E_{CRIT}, which is the intrinsic breakdown field of the

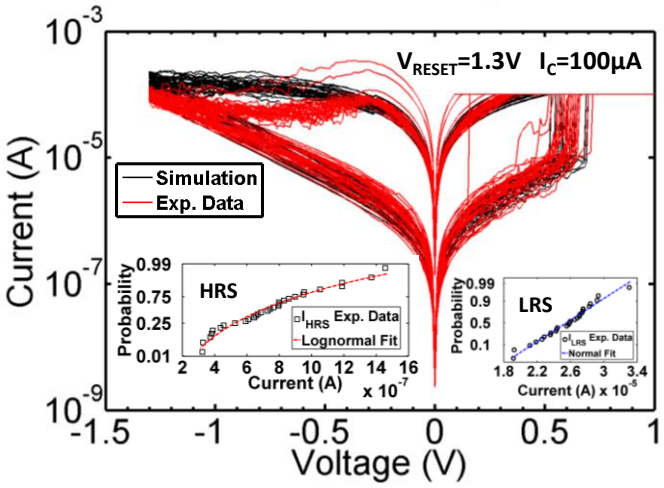

Figure 13. Experimental (red) and simulated (black) I-V curves of 30 consecutive full switching cycles for a 1µm^2 TiN/5nm-Ti/5nm-HfO_2/TiN device formed at I_C=100 µA and reset at V_{RESET}=1.3 V. T=25°C. Ramp rate=1V/s. Statistical variations of HRS and LRS resistance values and set/reset dynamics are correctly reproduced. The insets show the lognormal fitting of the HRS current and the normal fitting of the LRS current, both sensed at V_{READ}=100mV.

material (HfO_2), then the barrier effectively breaks down. Typical E_{CRIT} values for the sub-stoichiometric HfO_2 lie in the range of 4-6 MV/cm [24]. In this compact representation, the current during the set operation is calculated using (5) until the electric field reaches the critical value. In this circumstance the device resistance is set to the LRS value, i.e. R=R_{LRS}, and the current is calculated through Ohm's law (limited by the current compliance I_C, imposed during forming). The validity of the compact model is summarized in Figs. 10 to 12. Figure 10 a,b) shows the good matching between experimental (red curve) and simulated (dashed blue line) I-V curves capturing a complete switching cycle. Nicely, the compact model achieves a good agreement with the experimental data acquired from devices produced by different manufacturers, see Fig. 10 a,b). Figure 11 shows how the model can correctly reproduce the I-V characteristics of devices formed at different compliances (i.e. I_{C1}=100 µA and I_{C2}=200 µA). The model also correctly takes into account the effect of different reset conditions. This is shown in Fig. 12, where the simulated I-V curves of the reset operation performed with three different V_{RESET} values (i.e. V_{RESET1}=1.1V, V_{RESET2}=1.3V and V_{RESET3}=1.5V) are compared to the experimental data, showing a good agreement.

C. *Variability*

So far, the compact model is completely deterministic. This implies that the simulation of a given switching operation (e.g. reset) will always produce the same results, when starting from the same initial conditions. However, this is in contrast with the actual response of the device, which is largely stochastic [16-20]. Indeed, the physics of the switching operations show some intrinsic randomness, which determines a significant variation of the device characteristics (i.e. LRS and HRS resistance) over cycling. This is highlighted in Fig. 13, where the experimental I-V curves (red lines) of a device acquired during 30 consecutive full switching cycles (always

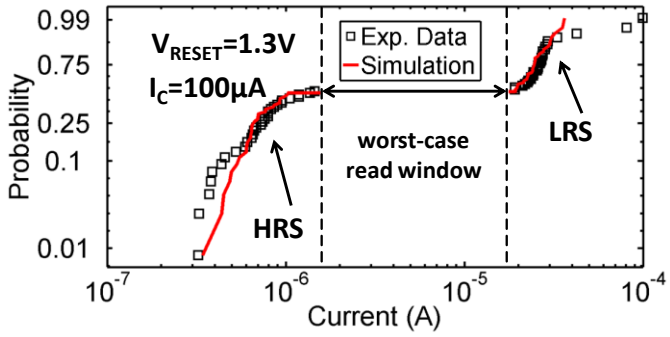

Figure 14. Experimental (black squares) and simulated (red solid line) probability plot of the absolute value of the current in both LRS and HRS, sensed at the reading voltage (V_{READ}=0.1 V), for a 1µm^2 TiN/5nm-Ti/5nm-HfO_2/TiN device formed at I_C=100 µA and reset at V_{RESET}=1.3 V (I-V curves are reported in Fig. 13). T=25°C. Ramp rate=1V/s. The model can reproduce the experimentally observed variability in both resistive states. The worst-case read window, including cycling variability, is highlighted.

performed in the same conditions) are reported. The significant variability results from the stochastic nature of the physical processes involved in the switching operations, i.e. diffusion and recombination of oxygen ions with vacancies in the CF and generation of O/Vo pairs, see Section II. Particularly, the events involved in the reset operation and assisting the growth of the dielectric barrier (recombination of O/Vo pairs) are due to the interaction between individual defects and can be considered as independent [24]. As a result, when reset is performed always in the same conditions, we may expect the barrier to show different thickness values over cycling. Particularly, we expect the barrier thickness to follow the normal distribution. The device resistance in HRS is linked to the barrier thickness by an exponential relation, eq. (3), which results in the lognormal distribution associated with the HRS resistance, see the inset of Fig. 13. These characteristics are included in the compact model by considering the barrier thickness as a gaussian distributed variable.

Likewise, the intrinsic randomness involved in the set operation results in slight variations of the CF properties over cycling. As a consequence, the LRS resistance over cycling is normally distributed. As such, R_{LRS} is also considered as a gaussian distributed variable, inset in Fig. 13. The compact model can correctly reproduce experimental cycle-to-cycle variations and both HRS and LRS resistance distributions, Figs. 13 and 14. Consequently, it can be used to estimate the worst-case read window, Fig. 14, i.e. a performance indicator which includes the detrimental effect of cycling variability. The small current fluctuations during set and reset operations due to stochastic generation and recombination of oxygen vacancies are neglected in this compact description.

D. *Parameters Calibration*

Besides many advantages, as the possibility of correctly reproducing variability, the compact model also displays ease of implementation. Indeed, it requires the calibration of a very small parameters set, the extraction of which involves only three I-V measurements: forming, set, and reset.

Indeed, S (i.e. the CF cross-section) can be estimated through eq. (1), provided the R_{LRS} value, which is measured at

Parameter	Value (R.T.)	Measurement
V_0	0.33V	HRS I-V
V_{INIT}*	-0.28V	Reset I-V
E_{CRIT}*	5.2MV/cm	Set I-V
E_A	0.05 – 0.08eV	R_{HRS} Arrhenius Plot
c_{xv}#	$1.1 \cdot 10^{-7}$A·nm·s^{-1}	Reset and HRS I-V
$c_{x,var}$	0.034	HRS I-V (cycling)
$c_{rset,var}$	0.08	LRS I-V (cycling)

*The value of this parameter may change with temperature
#This parameter was evaluated by applying a DC sweep reset with 500mV/s ramp rate

Table I. Parameters of the compact model and extracted values at room temperature.

V_{READ}=0.1 V right after the forming operation. V_{INIT} in (4) is extracted from the reset I-V curve as the voltage at which the current first departs from linearity more than 10%, hence modeling the voltage at which the barrier growth effectively begins (i.e. when O/Vo pairs start recombining, which first requires the O ions to be pushed down toward the bottom CF tip, see Section II) . c_{xv} in (4) is the parameter for best fitting of the reset operation I-V curves, Fig. 12. This parameter controls the barrier growth during reset and, as such, may also be used to include its temperature dependence [56-57]. This would require the estimation of the c_{xv} parameter at different temperatures, to include the temperature effect on the reset dynamics. V_0 in (5) is extracted by fitting the experimental HRS I-V curve, Fig. 12. E_A in (3) is extracted from the Arrhenius plot of the HRS resistance, Fig. 8. E_{CRIT} is an intrinsic material property, i.e. the critical field for dielectric breakdown. Its value can be also estimated by reproducing the I-V curves during set, Fig. 11: a higher value of E_{CRIT} causes the transition from HRS to LRS to occur at a higher voltage. This parameter could also be extracted at different temperatures to include the temperature dependence of the set process, if needed. Finally, cycling set and reset operations is required to estimate the parameters associated with variability features, Fig. 13. From HRS and LRS distributions (see insets in Fig. 13) we can extract $c_{x,var}$ and $c_{rset,var}$. In this framework, $c_{x,var}$ is the variance of the normal distribution associated with the lognormal R_{HRS} distribution. Correspondingly, $c_{rset,var}$ is the variance of the normal R_{LRS} distribution.

The parameter values extracted at RT and used to reproduce the whole experimental dataset in Figs. 10 to 14 are listed in Tab. I. Using these parameters values, the model correctly reproduces the switching dynamics as well as the experimental I-V cycling variability, matching the probability distributions of the currents measured in read conditions (V_{READ}=0.1 V) in both HRS and LRS. Figure 14 shows the comparison between the experimental and the simulated cumulative probability of the HRS and LRS currents. The excellent agreement in HRS is given by the powerful yet simple modeling of the stochastic variations of the barrier thickness, which reflects the physical randomness associated with the reset operation. The excellent agreement between measurements and simulations for devices manufactured with different process technologies and operated in different conditions confirms the validity of the proposed model. Notably, the E_{CRIT} value estimated from the set operation I-V curves, 5.2MV/cm – see Tab. I, is quantitatively in agreement with the theoretical predictions for sub-stoichiometric HfO_2, 4-6MV/cm [24], further strengthening the validity of the proposed approach.

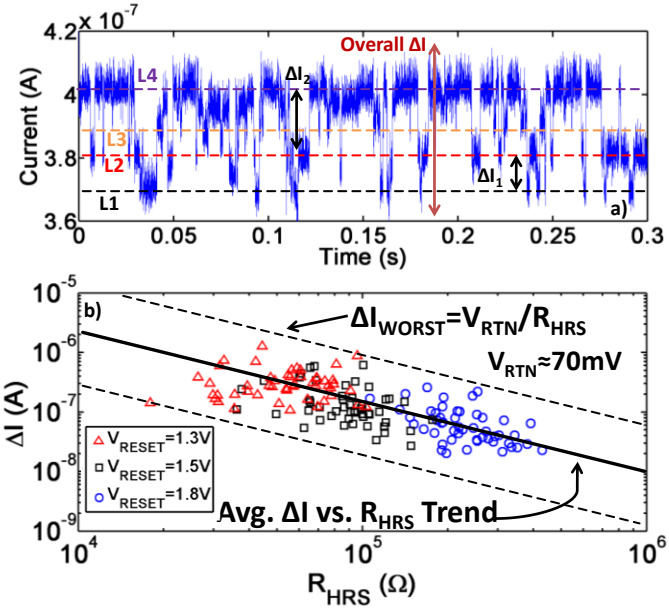

Figure 15. a) Experimental multi-level RTN trace as detected in an RRAM device in HRS. Four levels can be distinguished (L1 to L4). This signal can be identified as the summation of two independent two-level RTN fluctuations, whose amplitudes are ΔI_1 and ΔI_2. The overall noise amplitude is $\Delta I = \Delta I_1 + \Delta I_2$. b) Scatter plot of the ΔI (overall noise amplitude) vs. R_{HRS} for a device cycled 50 times in three different reset conditions. No point-to-point correlation exists between ΔI and R_{HRS} but only a statistical link. The average trend of ΔI with R_{HRS} is reported (black solid line). ΔI_{WORST} is the maximum noise amplitude at a given R_{HRS}. Its trend is highlighted by the upper dashed line.

IV. Random Telegraph Noise

Besides cycling variability, another major concern affecting RRAM behavior, especially during reading, is the Random Telegraph Noise (RTN) [13, 58-64]. RTN appears as sudden sharp and unpredictable fluctuations of the current between two (two-level RTN) or more (multi-level RTN) discrete levels, see Fig. 15a). In the case of RRAM devices, RTN mostly emerges during the read operation, reducing the effective readout margin of the device and potentially causing read failures [58-64]. The RTN disturb represents today one of the most significant limitations for the full exploitation of RRAM technology. It is hence important to understand its physical origin in order to control and predict its effects. Moreover, since RTN provides an additional source of variability, its effects must be also included in the compact model to achieve variability-aware and noise-resilient RRAM-based circuits design. Currently, the physical mechanism responsible for RTN fluctuations in RRAM is still under debate [58-64]. Nevertheless, there are evidences indicating that RTN results from the activation and de-activation of Vo defects assisting the TAT charge transport in the barrier [61-62]. A thorough investigation of the RTN phenomenon in these devices may be found in [62], which also provides a consistent physical discussion about the mechanisms possibly

triggering RTN.

The existence of RTN current fluctuations during the read operation has a detrimental effect on RRAM reliability. Indeed the amplitude of the overall noise fluctuation, ΔI, adds up on the cycling variability of the HRS and LRS resistance (or, equivalently, HRS and LRS current), determining a further reduction of the worst-case read window. As such, a reliable estimation of the worst-case read window should also consider the effect of the RTN fluctuations. Since the effects of cycling variability and RTN occur simultaneously, the eventual correlation between the two phenomena has to be investigated. Here we focus on the HRS, as the effect of RTN fluctuation is found to be critical only in this resistive state [61-62]. In order to study the correlation between RTN and the I_{HRS} cycling variability, we analyzed the RTN fluctuations (applying V_{READ}=50mV) and I_{HRS} (or, equivalently, R_{HRS}) at 50 consecutive switching cycles and in different reset conditions. Figure 15b) shows a scatter plot of the HRS resistance and the overall RTN current fluctuations (i.e. the difference between the maximum and the minimum value of the current in the RTN trace). Dissimilar ΔI values are found for comparable R_{HRS} values and vice versa, indicating that R_{HRS} (or I_{HRS}) variations due to cycling variability are not directly correlated to the RTN current fluctuations, Fig. 15b). However, R_{HRS} displays a statistical relation with the RTN fluctuations amplitude ΔI, see the trend in Fig. 15b). The maximum ΔI detected during RTN measurements, ΔI_{WORST} in Fig. 15b, depends directly on the R_{HRS} value and can be modeled as $\Delta I_{WORST}=V_{RTN}/R_{HRS}$, where $V_{RTN}\approx$70mV is the proportionality factor. Since R_{HRS} is measured applying a constant voltage (V_{READ}), ΔI_{WORST} can be converted to the worst-case resistance variation $\Delta R_{WORST}(R_{HRS})=V_{READ}/\Delta I_{WORST}=(V_{READ}/V_{RTN})\cdot R_{HRS}$.

The compact model does not allow reproducing the time evolution of RTN but can include safety margins to take into account the worst-case RTN effect. To do so, we consider the RTN-induced resistance variation as an extra variability source which sums up to cycling variability (being statistically independent from it). This further increases the R_{HRS} statistical spread in Figs. 13-14 (i.e. the variance of the distribution – $c_{x,var}$ in Tab. I). To quantify this effect we can modify the R_{HRS} distribution in Figs. 13-14 by including for each R_{HRS} value also the worst-case (maximum and minimum) RTN-induced resistance values, namely $R_{HRS}\pm\Delta R_{WORST}(R_{HRS})/2$. The factor of ½ arises from ΔR_{WORST} being centered on R_{HRS}. Evaluating the variance of the modified distribution gives back the new $c_{x,var}$ value which includes conservative margins accounting for the possible presence of RTN fluctuations.

V. Variability Control: Program-Verify Schemes

The comprehensive understanding of the RRAM devices developed here plays a key role in the identification of the fundamental limitations of this technology. The modeling approach used in this work naturally emphasizes the stochastic nature of many processes involved in the RRAM device operations, catching the intrinsic link between the physical mechanisms responsible for the electrical characteristics of these devices and the significant variability sources affecting them. This provides a very important feedback to device manufacturers for technology development and improvement. Indeed, while the superior features of these non-volatile memories [1-10] (e.g. switching speed, energy consumption, scaling, ease of implementation) look very attracting, the effects of variability issues pose significant challenges for the memory design. These issues may only partially be addressed by improvements in the manufacturing process and stack engineering [42-44]. Indeed the physics-based analysis proposed here successfully evidences how cycling variability and RTN result from physical mechanisms connected to the device operations [62]. As a result (and similarly to other NVM technologies, e.g. Flash) it is unlikely to achieve a tight control over variability, even in optimized cells with finely tuned stack properties. As such, program-verify algorithms, widely used in Flash memories, are required [65] to attain a good control over both cycling variability and RTN. In the following we show how the implementation of a refined program-verify scheme tailored on the physics of the device can be used to minimize the resistance dispersion in HRS, eventually allowing reliable RTN-resilient multi-bit operations, which may contribute to lower the implementation costs of this technology. To this point, the results of the physics-based analysis are essential to properly design the program-verify scheme. The underlying algorithm [66] exploits indeed the stochastic response of RRAM to the reset operation and the possibility of controlling the average HRS resistance through the reset voltage.

A. Variability Control

Figure 16 shows the probability distributions of both R_{LRS} and R_{HRS} over 50 pulsed switching cycles performed on a TiN/10nm-Ti/10nm-HfO_2/TiN device. Here we use 10µs wide pulses for set and reset operations. The three different R_{HRS} distributions are related to three different reset conditions, i.e. three different V_{RESET} values. As expected, $R_{LRS}\approx$10kΩ follows a tight normal distribution, consistently with the physical understanding of the device switching and with the outcomes shown in Figs. 13 and 14. Conversely, the log-normal R_{HRS} distribution significantly depends on V_{RESET}, Fig. 16, and shows a relevant dispersion. Moreover, these findings confirm the validity of the compact model also when dealing with pulsed switching, since the resistance distributions and the trend of the average resistance with the reset voltage, Fig. 16, are consistent with those found when using DC sweeps, Figs. 9-14, which could indicate that quasi steady-state approximations holds in this case. To unambiguously identify the device state after each operation, the device must exhibit a sufficiently high worst-case read window ($R_{HRS,MIN}/R_{LRS,MAX}$), which is lower than the median read window ($R_{HRS,50\%}/R_{LRS,50\%}$) due to the significant R_{HRS} dispersion, see the inset in Fig. 16. Moreover, the worst-case read window should also include the effect of RTN, which causes its further reduction. As long as the device is used to store a single bit (i.e. only two distinct resistive states are needed), reliable switching is easily achieved as the worst-case read window is always sufficiently large, see the inset in Fig. 16. For the sake

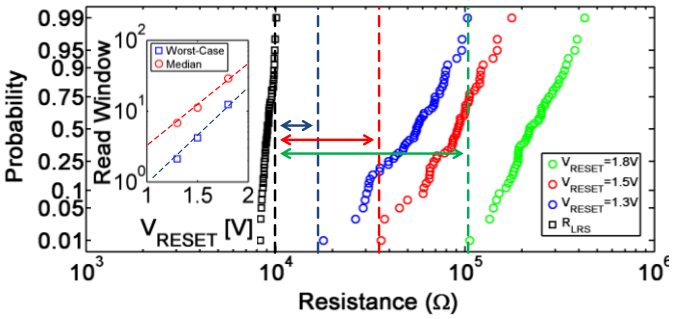

Figure 16. R_{LRS} and R_{HRS} distributions for three different V_{RESET} obtained on a 0.6x0.6 μm^2 TiN/10nm-Ti/10nm-HfO_2/TiN device formed with I_C=100μA. The colored and black vertical dashed lines represent the minimum of each R_{HRS} distribution and the maximum R_{LRS}, respectively. The worst-case read window is indicated by an arrow for each V_{RESET}. Since the three R_{HRS} distributions overlap, no multi-bit operation is allowed. In the inset, the worst-case (blue squares) and the median (red circles) read window vs. V_{RESET} is reported. The R_{HRS} dispersion causes the worst-case read window to be significantly lower than the median value.

of clarity, the worst-case read window shown in Fig. 16 does not include the effect of RTN. However, since the average value of R_{HRS} can be modulated by employing different reset conditions (i.e. different V_{RESET}), we may define more than one HRS, which may eventually enable the storage of more than one bit in a single cell [65-67]. Regrettably, the R_{HRS} distributions associated with the three different V_{RESET} in Fig. 16 are largely overlapped, impeding multi-bit operations. As such, multi-bit storage can only be achieved by providing a tight control over the R_{HRS} dispersion. To this point, the application of program-verify schemes [65-66] may be utterly beneficial. For instance, program-verify schemes are largely used in Flash memories [65] to guarantee adequate reliability and to achieve optimized control over variability. The concept of the program-verify approach relies on performing a write operation (program), to then check if the targeted resistance level is reached (verify) [65-67]. Based on the result of the verify step, the underlying algorithm may either stop its execution (in case of success) or re-iterate the procedure (in the opposite case). Usually, an exit condition is considered (i.e. a maximum number of algorithm iterations) to keep the total execution time within reasonable limits [65-67]. So, the algorithm will fail if is not able to succeed before the exit condition.

As such, the design of an efficient and optimized program-verify scheme has to be tailored on the device physics to guarantee high success rate, minimum total execution time, and optimal control over the R_{HRS} dispersion.

B. Dispersion-Aware Program-Verify Scheme

Recently, we proposed the Dispersion-Aware Program-Verify (DAPV) scheme [66] which exploits the stochastic response of RRAM to the reset operation and offers the following advantages:

1) A single iteration of the DAPV consists only in the sequence of a set, reset, and read operation, plus at most five comparison operations. This represents a limited computational overhead and increase of cost on silicon, compared to classical program-verify schemes needing at least three comparison operations [66]. Each of these operations can be performed in few nanoseconds, which makes the algorithm very fast.
2) The algorithm performs a real-time estimation of R_{HRS} dispersion and automatically changes the reset conditions to minimize the total execution time.
3) As such, the algorithm is also able to track possible device degradation over time (changes in R_{HRS} distributions), which is experimentally observed in these devices.
4) The user can define an arbitrary number of different HRS to be targeted, each characterized by a target resistance span (R_{MIN} to R_{MAX}). This makes the R_{HRS} dispersion user-defined.
5) The user can trade-off the average total execution time with the number of different HRS defined (i.e. number of stored bits vs. performance).
6) DAPV offers higher success rate and quicker convergence compared to classical schemes.

The details of this novel scheme are reported elsewhere [66] and will not be discussed here. However, in this work we report some results of the application of this scheme to RRAM devices and provide previously unreported additional details. Particularly, we apply the DAPV to the same device used to derive the distributions in Fig. 16. To this point we define three different HRS, aiming at achieving reliable two-bits/cell operations (three HRS and one LRS). For each HRS we define the corresponding target resistance span:

1) R_{MIN}=40kΩ, R_{MAX}=60kΩ
2) R_{MIN}=70kΩ, R_{MAX}=100kΩ
3) R_{MIN}=200kΩ, R_{MAX}=300kΩ

At this point we perform 100 DAPV runs for each targeted HRS. The reset, set, and read operations are always performed in pulsed mode, by applying 10μs wide pulses. To this point it is noteworthy that the algorithm performances are not significantly affected by the pulse width. Here we choose a 10μs pulse width to minimize measurement instrumentation noise but a much shorter pulse may be used to switch the device. While the reset voltage is dynamically modulated by the DAPV scheme, the voltages applied during the set pulse (which has to be higher than the typical V_{SET} value) and the read pulse are always 2V and 0.1V, respectively. Moreover, we set the maximum number of iterations to 20. The resulting R_{HRS} distributions are reported in Fig. 17, showing the superior control of the R_{HRS} dispersion. Table II reports information about the performance of the DAPV. The algorithm always succeeds in less than 10 iterations, confirming the possibility of reliable two-bits/cell operations. However, this portrait still does not include the detrimental effect of RTN. Indeed the margins between neighboring resistance distributions in Fig. 17, represented by dark-shaded bands, are derived by considering only the resistance dispersion (i.e. cycling variability). However, the refined control over cycling variability also results in a better control

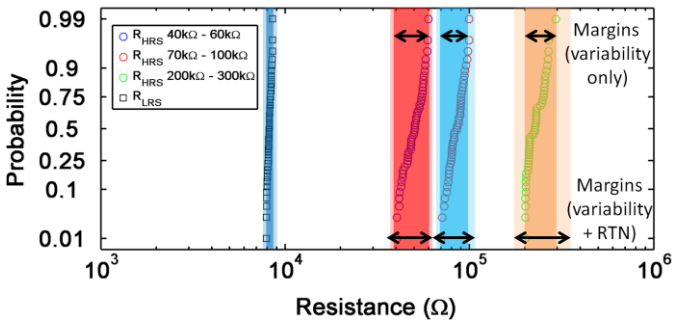

Figure 17. R_{LRS} and three R_{HRS} distributions as obtained with DAPV on the same device analyzed in Fig. 16. For each target R span the algorithm is run 100 times with no failures, obtaining superior control over variability. Each distribution is surrounded by a dark-shaded band (which highlights the effect of cycling variability only – it runs from R_{MIN} to R_{MAX}) and a light-shaded band (which also includes the detrimental effect of the worst-case RTN). Target R spans can be chosen by the user to achieve reliable multi-bit operations with RTN-resilience.

Target R Span	Avg. Iter.	Median Iter.	Max Iter.	Min Iter.
40kΩ - 60kΩ	2.8	1.5	9	1
70kΩ - 100kΩ	2.88	2	9	2
200kΩ - 300kΩ	2.42	2	9	1

Table II. Performance of the DAPV. Each target resistance (R) span is programmed 100 times by means of the DAPV. The average, median, maximum, and minimum number of iterations performed by the DAPV to successfully program the cell is reported for each target resistance span. DAPV always succeeds in less than 10 iterations.

over the RTN detrimental effect. In fact, the tight control of the target resistances span boundaries (i.e. R_{MIN} and R_{MAX} for each distribution) implies a tight control also on the worst-case ΔI, see Fig. 15b). As such, this value can be estimated beforehand and can be hence taken into account while defining the target resistance spans. In Fig. 17 we show how including the worst-case ΔI due to RTN results in a reduction in the worst-case read window between neighboring distributions (which are enlarged, see the light-shaded bands). However, the target resistance spans chosen in this study are sufficiently apart from each other to guarantee RTN-resilience Indeed, the worst-case read window, including the worst-case RTN effect, is always larger than unity, which guarantees reliable and fast two-bits/cell operations with superior variability control and RTN resilience.

VI. Conclusions

In this paper we thoroughly investigated the characteristics of the TiN/Ti/HfO_2/TiN RRAM device. Initially, we explored the physical mechanisms involved in the device operations by considering the material-specific atomistic features of the HfO_2. Self-consistent simulations were used to obtain a full comprehension of the device operations from the physics standpoint. Device operations involve several charge and ion transport phenomena, in addition to structural modifications of the device at the atomistic level which assist the device switching. The comprehensive physical understanding allowed capturing the reasons for variability, which is due to the intrinsic randomness in the physical mechanisms involved in the device switching. By connecting the electrical response of the device to the atomistic material properties, we devised also a physics-based compact model. It correctly describes the device behavior in different operating conditions, including also the effects of variability. Moreover, we also studied the statistical connections between RTN and cycling variability. The existence of a statistical link between the programmed resistance level and the worst-case RTN read disturb allowed including RTN effects in the compact model. Moreover, we demonstrated that superior control over variability and noise may be obtained by relying on an advanced programming scheme tailored on the device physics. Its implementation allowed achieving reliable and RTN-resilient two-bits/cell operations.

References

[1] Y.-S. Chen et al., "Highly scalable hafnium oxide memory with improvements of resistive distribution and read disturb immunity", Proc. of IEEE Electron Devices Meeting (IEDM), pp.1,4, 7-9 Dec. 2009

[2] C. Nauenheim et al., "Nano-Crossbar Arrays for Nonvolatile Resistive RAM (RRAM) Applications", NANO '08 8th IEEE Conference on Nanotechnology, pp.464,467, 18-21 Aug. 2008.

[3] A. Chen, Ionic Memory Technology, in Solid State Electrochemistry II: Electrodes, Interfaces and Ceramic Membranes, First Edition (ed V. V. Kharton), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany. 2011. doi: 10.1002/9783527635566.ch1.

[4] B. Govoreanu et al., «10x10nm2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation», in IEDM Tech. Dig., Washington, DC, USA, 2011.

[5] S. L. Lee et al., «Multi-level switching of triple-layered TaOx RRAM with excellent reliability for storage class memory», Symposium on VLSI Technology (VLSIT), pp. 71 - 72, 2012.

[6] H. Y. Lee et al., «Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO_2 based RRAM», in IEDM Tech. Dig., San Francisco, CA, USA, 2008.

[7] B. Butcher et al., «High endurance performance of 1T1R HfOx based RRAM at low (<20μA) operative current and elevated (150°C) temperature», IEEE International Integrated Reliability Workshop Final Report (IRW), pp. 146 - 150, 2011.

[8] C. W. Hsu et al., "Self-rectifying bipolar TaO_x/TiO_2 RRAM with superior endurance over 10^{12} cycles for 3D high-density storage-class memory", Symposium on VLSI Technology, pp. 166 - 167, 2013.

[9] S. Deora et al., "Statistical assessment of endurance degradation in high and low resistive states of the HfO_2-based RRAM", IEEE International Reliability Physics Symposium, pp.MY.2.1-MY.2.5, 14-18 April 2013.

[10] H.-S.P. Wong et al., "Metal–Oxide RRAM", Proceedings of the IEEE, vol.100, no.6, pp.1951,1970, June 2012.

[11] S. Balatti, S. Ambrogio, D. Ielmini, D. Gilmer, «Variability and failure of set process in HfO2 RRAM», 5th IEEE International Memory Workshop (IMW), pp. 38 - 41, 2013.

[12] A. Fantini et al., «Intrinsic switching variability in HfO2 RRAM», 5th IEEE International Memory Workshop (IMW), pp. 30 - 33, 2013.

[13] F. Puglisi, L. Larcher, P. Pavan, A. Padovani, G. Bersuker, «Instability of HfO2 RRAM devices: Comparing RTN and cycling variability», Proceedings of the IEEE International Reliability Physics Symposium, pp. MY5.1 - MY5.5, 2014.

[14] N. Raghavan et al., «Stochastic variability of vacancy filament configuration in ultra-thin dielectric RRAM and its impact on OFF-state reliability», Proceedings of the IEEE Electron Devices Meeting (IEDM), pp. 21.1.1 - 21.1.4, 2013.

[15] B. Traore et al., «Impact of electrode nature on the filament formation and variability in HfO2 RRAM», Proceedings of the IEEE International Reliability Physics Symposium, pp. 5E.2.1 - 5E.2.5, 2014.

[16] L. Vandelli et al., "Modeling of the Forming Operation in HfO2-Based Resistive Switching Memories", Proc. IEEE Int. Memory Workshop, Monterey (CA), USA, May 22-25, 2011, pp. 119-122.

[17] S. Yu, R. Jeyasingh, Y. Wu, H. S. Wong, «Understanding the conduction and switching mechanism of metal oxide RRAM through low frequency noise and AC conductance measurement and analysis», in IEDM Tech. Dig., Washington, DC, USA, 2011.

[18] X. Guan, S. Yu, H.-S. P. Wong, «A SPICE Compact Model of Metal Oxide Resistive Switching Memory With Variations», IEEE Electron Device Letters, vol. 33, n. 10, pp. 1405 - 1407, 2012.

[19] B. Chen et al., «A compact model of resistive switching devices», IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT), pp. 1829 - 1831, 2010.

[20] M. Bocquet, D. Deleruyelle, H. Aziza, C. Muller, J.-M. Portal, «Compact modeling solutions for OxRAM memories», IEEE Faible Tension Faible Consommation (FTFC), 2013.

[21] W. C. Shen et al., "Memristor Logic Operation Gate With Share Contact RRAM Cell", IEEE Electron Device Letters, vol.32, no.12, pp.1650,1652, Dec. 2011.

[22] A. Singha et al., «Analog memristive time dependent learning using discrete nanoscale RRAM devices», International Joint Conference on Neural Networks (IJCNN), pp. 2248 - 2255, 2014.

[23] A. Chen, «Utilizing the Variability of Resistive Random Access Memory to Implement Reconfigurable Physical Unclonable Functions», IEEE Electron Device Letters, vol. 36, n. 2, pp. 138 - 140, 2015.

[24] G. Bersuker et al., «Metal oxide resistive memory switching mechanism based on conductive filament microscopic properties», Journal of Applied Physics, vol. 110, p. 124518, 2010.

[25] G. Bersuker et al., «Mechanism of electron trapping and characteristics of traps in HfO2 gate stacks», IEEE Trans. Device Mater. Rel., vol. 7, n. 1, pp. 138-145, 2007.

[26] S. Larentis et al., "Resistive Switching by Voltage-Driven Ion Migration in Bipolar RRAM—Part II: Modeling", IEEE Transactions on Electron Devices, vol.59, no.9, pp. 2468-2475, Sept. 2012.

[27] R. Degraeve et al., "Hourglass concept for RRAM: A dynamic and statistical device model", 21st IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA), 2014, pp. 245,249, June 30 2014-July 4 2014.

[28] N. Raghavan et al., "Monte Carlo model of reset stochastics and failure rate estimation of read disturb mechanism in HfOx RRAM", IEEE International Reliability Physics Symposium (IRPS), 2015, pp. 5B.2.1,5B.2.9, 19-23 April 2015.

[29] B. Traore et al., "On the forming-free operation of HfOx based RRAM devices: Experiments and ab initio calculations", Proceedings of the European Solid-State Device Research Conference (ESSDERC), pp. 170-173, 16-20 Sept. 2013.

[30] Z. Fang et al., "$HfO_x/TiO_x/HfO_x/TiO_x$ Multilayer-Based Forming-Free RRAM Devices With Excellent Uniformity", IEEE Electron Device Letters, vol.32, no.4, pp.566,568, April 2011.

[31] K. Young-Fisher et al., «Leakage Current-Forming Voltage Relation and Oxygen Gettering in HfOx RRAM Devices», IEEE Electron Device Letters, vol. 34, n. 6, p. 750, 2013.

[32] D. Walczyk et al., "Resistive switching characteristics of CMOS embedded Hf02-based 1T1R cells", Microelectron. Eng., vol. 88, no. 7, pp.1133-1135, Jul. 2011.

[33] http://www.mdlab-software.com

[34] L. Vandelli et al., "A physical model of the temperature dependence of the current through SiO_2/HfO_2 Stacks", IEEE Trans. Electron Devices, Vol. 58(8), pp. 2878-2887, 2011.

[35] M. Zhang et al., "Unification of three multiphonon trap-assisted tunneling mechanisms", Journal of Applied Physics, Volume 110, Issue 11, pp. 114108-114108-11 (2011).

[36] N. Capron et al., «Migration of oxygen vacancy in HfO2 and across the HfO2/SiO2 interface: A first-principles investigation», Appl. Phys. Lett., vol. 91, p. 192905, 2007.

[37] A. S. Foster et al., «Vacancy and interstitial defects in hafnia», Phys. Rev. B, Condens. Matter, vol. 65, n. 17, p. 174117, May 2002.

[38] A. S. Foster et al., «Mechanism of Interstitial Oxygen Diffusion in Hafnia», Phys. Rev. Lett. 89, vol. 89, n. 22, p. 225901, Nov. 2002

[39] G. Bersuker et al., «,Mechanism of high-k dielectric-induced breakdown of interfacial SiO2 layer», in IEEE Int. Rel. Phys. Symp., 2010.

[40] O. Ghobar et al., «Defects in the interfacial layer of SiO2-HfO2 gate stacks: depth distribution and identification», in Proc. IIRW, 2007.

[41] O. Blank et al., «A model for multistep trap-assisted tunneling in high-k dielectrics», J. Appl. Phys., vol. 97, n. 4, p. 044147, 2005.

[42] Y. Y. Chen et al., "Endurance/retention trade-off on HfO_2/metal cap 1T1R bipolar RRAM," IEEE Trans. Electron Devices, vol. 60, no. 3, pp. 1114–1121, Mar. 2013.

[43] C. Walczyk et al., «On the role of Ti adlayers for resistive switching in HfO_2-based metal-insulator-metal structures: Top versus bottom electrode integration», J. Vac. Sci. Technol. B, vol. 29, p. 01AD02, Jan/Feb 2011.

[44] Z. Fang et al., «The Role of Ti Capping Layer in HfO_x-Based RRAM Devices», IEEE Electr. Device Letters, vol. 35, n. 9, pp. 912-914, 2014.

[45] J. McPherson, J. Y. Kim, A. Shanware, H. Mogul, «Termochemical description of dielectric breakdown in high dielectric constant materials», Appl. Phys. Lett., vol. 82, n. 13, pp. 2121-2123, 2003.

[46] G. Broglia et al., «Molecular dynamics simulation of amorphous HfO 2 for resistive RAM applications», Modelling and Simulation in Materials Science and Engineering, vol. 22, n. 6, p. 065006, 2014.

[47] D. Muñoz Ramo et al., "Spectroscopic properties of oxygen vacancies in monoclinic HfO_2 calculated with periodic and embedded cluster density functional theory", Phys. Rev. B, Vol. 75, 2007, p. 205336

[48] R. Landauer, «Spatial Variation of Currents and Fields Due to Localized Scatterers in Metallic Conduction», IBM Journal of Research and Development, vol. 1, p. 223, 1957.

[49] J. Robertson, R. Gillen, «Defect densities inside the conductive filament of RRAMs», Microelectronic Engineering, vol. 109, pp. 208-210, 2013.

[50] Zhang et al., "Statistical analysis of retention behavior and lifetime prediction of HfOx-based RRAM", IEEE International Reliability Physics Symposium (IRPS), pp.MY.8.1-MY.8.5, 10-14 April 2011.

[51] F. M. Puglisi et al., «An empirical model for RRAM resistance in low- and high-resistance states», IEEE Electron Device Letters, vol. 34, n. 3, pp. 387-389, 2013.

[52] S. Privitera et al., "Conductive filament structure in HfO_2 resistive switching memory devices", Solid-State Electronics, vol. 111, Sept. 2015, pp. 161-165, ISSN 0038-1101.

[53] D. Veksler, et al., «Random telegraph noise (RTN) in scaled RRAM devices», Proceedings of the IEEE International Reliability Physics Symposium, pp. 10.1-10.4, 2013.

[54] D. Ielmini, F. Nardi, C. Cagli, "Universal Reset Characteristics of Unipolar and Bipolar Metal-Oxide RRAM," IEEE Transactions on Electron Devices, vol.58, no.10, pp.3246,3253, Oct. 2011.

[55] F. M. Puglisi et al., «A Compact Model of Hafnium-Oxide-Based Resistive Random Access Memory», Proceedings of the IEEE International Conference on IC Design and Technology (ICICDT), pp. 85 - 88, 2013.

[56] F. M. Puglisi, A. Qafa, P. Pavan, "Temperature Impact on the Reset Operation in HfO2 RRAM", IEEE Electron Device Letters, vol. 36, no. 3, pp. 244-246, Mar. 2015.

[57] R. Fang et al., "Low-Temperature Characteristics of HfOx-Based Resistive Random Access Memory", IEEE Electron Device Letters, vol.36, no.6, pp.567,569, June 2015.

[58] F. M. Puglisi et al., «Random Telegraph Signal Noise Properties of HfOx RRAM in High Resistive State», in Proc. of ESSDERC, pp.274-277, 17-21 Sept. 2012.

[59] F. M. Puglisi, P. Pavan, «RTN analysis with FHMM as a tool for multi-trap characterization in HfOX RRAM», IEEE International Conference of Electron Devices and Solid-State Circuits (EDSSC), pp. 1 - 2, 2013.

[60] N. Raghavan et al., "RTN insight to filamentary instability and disturb immunity in ultra-low power switching HfOx and AlOx RRAM", Symposium on VLSI Technology, pp. T164 - T165, 2013.

[61] F. M. Puglisi et al. «RTS noise characterization of HfOx RRAM in high resistive state», Solid-State Electronics, vol. 84, pp. 160 - 166, 2013.

[62] F. M. Puglisi et al., "A Complete Statistical Investigation of RTN in HfO2-based RRAM in HRS", IEEE Transactions on Electron Devices, vol.62, no.8, pp.2606-2613, Aug. 2015.

[63] F. M. Puglisi, P. Pavan, "Factorial Hidden Markov Model analysis of Random Telegraph Noise in Resistive Random Access Memories", ECTI Transactions on Electrical Engineering, Electronics, and Communications, vol. 12, no. 1, pp. 24-29, 2014.

[64] S. Ambrogio et al., «Statistical Fluctuations in HfOx Resistive-Switching Memory: Part II—Random Telegraph Noise», IEEE Transactions on Electron Devices, vol. 61, n. 8, pp. 2920 - 2927, 2014.

[65] S. I. Park, D. Shin, E. G. Han, "Adaptive program verify scheme for improving NAND flash memory performance and lifespan", IEEE Asian Solid State Circuits Conference 2012, pp. 57-60, 12-14 Nov. 2012.

[66] F. M. Puglisi, C. Wenger, and P. Pavan, "A Novel Program-Verify Algorithm for Multi-bit Operation in HfO_2 RRAM", IEEE Electron Device Letters, in press.

[67] A. Emara et al., "Differential 1T2M memristor memory cell for single/multi-bit RRAM modules", 6th Computer Science and Electronic Engineering Conference (CEEC), pp.69,72, 25-26 Sept. 2014.