HFO$_2$-BASED RESISTANCE SWITCHING NON-VOLATILE RANDOM ACCESS MEMORY:
LOW POWER OPERATION AND REDUCED VARIABILITY

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Asad Kalantarian
December 2015
I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Yoshio Nishi, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

S Wong

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Gennadi Bersuker

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.
Abstract

We have entered an era of “Big Data” in which the amounts of data being generated and stored is rapidly exceeding exascale. Information collected by various sensors, user data, personal information, and internet all have led to this rapid increase of stored information. In fact, not only has the amount of data increased drastically, but the nature of this data has changed as well. Computation is now required to revolve around this new data and traditional algorithms of predicting more frequently accessible data are now proving less useful. These changes have made memory capacity, performance, and reliability improvement more critical than ever.

The prospect of finding a Storage Class Memory (SCM) and its potential revolutionary impact on computer architecture has driven many researchers to investigate various alternatives to current FLASH based memory. As such, in this data-centric era efforts are being made to introduce emerging memory solutions that may bring about new level in the current memory hierarchy.

Resistance-change based random access memory (RRAM) based on Transition Metal Oxides (TMOs), whose operation is based on the change in resistivity due to
the formation of a conductive filament in the oxide material, has attracted attention in recent years because of its potential for high density, high speed, and good retention as a novel nonvolatile memory. However, achieving low power operation and high device-to-device uniformity in the cell resistance states are the major challenges for practical applications of RRAM technology. While certain progress has been made in understanding the switching mechanism of TMO memory devices, lack of precise control of the filament formation, perceived to be a random process, introduces variability into the switching characteristics of this class of devices, hindering further progress. In this thesis we address these major issues in HfO$_2$-based RRAM devices.

In this dissertation we address the problem variability and reducing switching power by proposing a constant voltage forming (CVF) method. The method is shown to increase the resistances of the low resistance and high resistance states while reducing their variability. By forcing the forming in all devices to occur at the same predefined voltage, the CVF method is demonstrated to eliminate a major cause of the device-to-device variation associated with the randomness of the forming voltage values. Moreover, both experiments and simulations show that CVF at lower voltages suppresses the parasitic overshoot current, resulting in a more controlled and smaller filament cross-section and lower operation currents.

Further, electrical and physical characterization of the HfO$_2$ RRAM devices is demonstrated, presenting physical evidence for a filament in HfO$_2$ RRAM using scanning transmission electron microscopy. A ‘deep reset’ phenomenon is observed electrically for ultra-short pulses at larger voltages otherwise not observable using longer pulse times. The interplay of breakdown and recombination forces for hafnium and
oxygen ions is suggested to explain this observed phenomenon.

Finally, we present a detailed model for the reset process kinetics in HfO$_2$-based RRAM describing the transition between low and high resistance states at the atomic level. Based on the filament characteristics as observed by TEM, the kinetics of the reset operation is then simulated using our developed Kinetic Monte Carlo (KMC) method incorporating ab-initio calculated microscopic characteristics of the oxygen ions in hafnia. Temperature and field driven oxygen diffusion in the oxide surrounding the filament is shown to provide the needed supply of oxygen to re-oxidize the tip of the filament and switch the device back to the High Resistance State (HRS). Using this developed model and KMC simulator, the dependence of reset process on surrounding interstitial oxygen ion concentration, reset pulse width, and reset pulse height is studied and the variability of the resulting HRS state is simulated.
Acknowledgment

All thanks and praise is due to God for providing for me the blessings, means and guidance to reach where I am. I thank Him for the opportunities He has set forth for me and pray that I can appreciate all that He has given me as it is truly deserved.

My PhD years at Stanford University have truly been an exceptional time in my life. With all the ups and downs have been wealth of experience and learning that I have gained and this has only been possible through the support and help of everyone around me of whom I am very grateful and I know that words are not capable of expressing the deepest gratitude and appreciation that I have towards them.

In particular, I would like to begin by thanking my advisor, Professor Yoshio Nishi, for being there from the very start, and for his support, advice, and flexibility with me in all the situations and circumstances that I have encountered during my PhD years. I also thank Professor Nishi for the opportunity to spend time at SEMATECH research consortium where I was able to perform industry grade research and connect with the best scientists and industry researchers in the field. As such, I thank my mentor at SEMATECH, Dr. Gennadi Bersuker and expert in reliability physics with
intuition that is beyond comprehension. I thank Dr. Bersuker from whom I have been able to benefit from his knowledge, company and valuable time which he generously gave to me.

I would like to extend my special thanks to my PhD oral exam committee members Prof. Nishi (advisor), Dr. Bersuker (SEMATECH Fellow), Prof. Simon Wong, and Prof. Evan Reed (the chair), Prof. Eric Pop. I am especially indebted to Prof. Simon Wong for agreeing to be my dissertation reader as well and his continued support.

I also would like to thank Dr. Raj Jammy, Dr. Paul Kirsch, and Dr. David Gilmer for all their help and guidance during my time at SEMATECH. Many thanks also to my colleagues at SEMATECH, whom are too many to mention, but in particular Dr. Brian Butcher for going through our PhD years together and our close collaboration and encouragement.

My gratitude goes out to all my friends and colleagues in the Electrical Engineering Department at Stanford and especial Professor Nishi’s nano-electronics group, and in particular Dan Duncan for all the adventures we had exploring Europe with such passion and energy during our trip to ST Microelectronics. I would like to thank all of my dear friends, especially those who would attend Dua Kumayl weekly, for filling me with joy and happiness with their presence and company — namely my brothers, Dr. Mazhareddin Taghivand, Dr. Morteza Ibrahimi, Dr. Hamid Bazargan, and especially may dear roommate Dr. Mohammad Bazargan.

I thank Allah for giving me my dearest parents, my dear sisters and my dear
brother Hamid all of whose support, guidance, encouragement, understanding, kindness and unconditional love have helped make me the person I am today. I thank my mom for her prayers and for wishing me the best always, even though destiny didn’t allow her to live to be with us physically. Last but not least, I give my deepest and sincere appreciation to my father Dr. Enayatollah Kalantarian for his patience with me all the time and for guiding me step by step with love, enduring the difficulty of me being away from home. I know I can never repay him for all that he has done for me and I pray that Allah repays him infinitely more than what he deserves both in this world and in the hereafter.
Dedication

To My Dear Father Dr. Enayatollah Kalantarian and My Dear Mom
# Contents

Abstract iv

Acknowledgment vii

Dedication x

1 Introduction 1

1.1 Evolution of Non Volatile Semiconductor Memory Technology 1

1.2 Storage Class Memory and the Data-Centric Era 4

1.3 RRAM: The Resistive Switching Phenomenon in Oxides 7

1.4 Operational Characteristics of HfO$_2$-Based RRAM 9

1.5 Thesis Overview 12
2 Controlling Uniformity of RRAM Characteristics

2.1 Sources of Conductive Filament Variability ............................................. 16

2.2 Caused by Compliance Current Overshoot .............................................. 19

2.2.1 Modeling the Impact of $C_P$ on Overshoot Current ............................ 22

2.3 Controlling the Filament Formation Process .......................................... 25

2.3.1 Simulating the Impact of Forming Conditions ..................................... 26

2.3.2 Measuring the Impact of Forming Conditions .................................... 30

2.4 Oxygen Ion Diffusion during the Forming Process ................................. 40

3 Physical and Electrical Characterization .................................................. 45

3.1 Observation of Conical Filament in HfO$_2$ RRAM ................................. 47

3.2 Kinetics of RRAM Switching using Ultra-Short Pulse Characterization ...... 51

3.3 Voltage vs. Time Trade-off and Deep Reset Phenomenon ....................... 58

4 Kinetics Model of the Reset Process ......................................................... 63

4.1 Modeling the Forming and Switching Processes .................................... 64

4.1.1 The Role of Grain Boundaries ......................................................... 64

4.1.2 Mechanism of Forming Process ...................................................... 65
4.1.3 The Reset Process ........................................... 69

4.2 3D Time Dependent Monte Carlo (TDMC) Simulation of Switching . 71

4.2.1 The Diffusion Process ...................................... 71

4.2.2 Local Temperature and Electrics Field Calculation ............ 74

4.2.3 Barrier Creation and Tunneling Mechanism ..................... 81

4.2.4 TDMC Simulation Flow ...................................... 85

4.3 Simulation of Reset Conditions .................................. 89

4.3.1 Pulse Length ............................................... 90

4.3.2 Pulse Amplitude ........................................... 91

4.3.3 Initial Ion Distribution .................................... 94

4.3.4 Variability .................................................. 97

5 Conclusions ................................................................ 100

5.1 Summary of Contributions ........................................ 100

5.2 Outlook and Future Work ......................................... 102
List of Tables
List of Figures

1.1 The pyramid-like memory hierarchy today. Current architecture is based on SRAM, DRAM, Flash and hard drive mass storage. The gap between memory and storage can be filled with SCM type memories. 6

1.2 Sketch of the Metal-Insulator-Metal (MIM) structure of HfO$_2$-based RRAM. .......................... 10

1.3 Complete switching curve of a typical HfO$_2$-based RRAM device under forming, reset and set voltage sweeps. ........................................ 11

1.4 Sketch of the forming and reset models in which oxygen ions are generated during forming, and interstitial oxygen ions oxidize the tip of the filament during reset. ........................................ 12

2.1 Examples of devices with (left) overshoot during forming as manifested by the $I_{Max}$ value being above the compliance limit, and (right) without overshoot. Maintaining a compliance limit during forming results in higher LRS and HRS resistances. .......................... 18
2.2 Usual distribution of forming voltages across different devices during a typical forming sweep. ........................................... 19

2.3 Dependence of the maximum of the 1\textsuperscript{st} reset current ($I_{Max}$) on the forming voltage, $V_F$. ................................................................. 20

2.4 Schematic of the measurement setup. The transient current $I_{Cp} = C_P \frac{dV_1}{dt}$ (charging of the parasitic capacitance $C_P$) is estimated using the $\frac{dV_1}{dt}$ values as measured by the oscilloscope (at node $V_1$). .......... 21

2.5 (a) Examples of the overshoot currents and the corresponding $I_{Max}$ values - $I_{Max}$ (peak current) is measured in the first reset after forming- (the “intrinsic” variation illustrated in Fig 2.3) for 4 devices during the forming event under the identical CVF condition of 1.7V with the 100\,µA compliance. (b) The 1\textsuperscript{st} reset (negative voltage) and set (positive voltage) of the corresponding devices. ......................... 22

2.6 Simulated final breakdown phase of forming approximated by an exponential time dependence. ................................................. 23

2.7 An example of the modeling (broken line) of the overshoot current (device 2 in Fig. 2.5) through RRAM. The parasitic capacitance is estimated to be around 1nF (using an external transistor in the 1T1R setup). ................................................................. 24
2.8 Simulated currents through RRAM during breakdown (BD) phase for three BD transient times $t_r = 2e-7s, 2e-8s, 2e-9s$ for the conditions as in Fig. 2.7 ................................................................. 25

2.9 Parasitic capacitance impact on overshoot current, simulated for $C_p = 1nF$ and 1pF using the conditions as in Fig. 2.7 ................................................................. 26

2.10 An example of the TDDB simulations using the forming model: CVF at 2V, T=225C, 20 devices with randomly selected initial trap distributions. For each device (each line), the current transient rate $\frac{dI}{dt}$ is defined by the time required for the current to grow from the initial (arbitrary) level of 1µA to the final BD (compliance) current. ...... 28

2.11 Transient rate ($\frac{dI}{dt} \sim \frac{-dR}{dt}$) calculated from the TDDB simulations in Fig. 2.10. Each point represents the average of 20 simulated devices at $T = 25^\circ C$ at the given CVF voltage. .................................................. 29

2.12 Simulated $\frac{dI}{dt}$ dependence on ambient temperature at 2V CVS, as well as number of traps needed in the filament to reach the forming compliance; a weaker dependence than on the forming voltage is observed. 30

2.13 Initial Leakage of randomly selected group of fresh devices before any forming. ................................................................. 31

2.14 Cumulative distribution of $I_{Max}$ for devices formed by different CVFs under 100µA compliance. The values of $I_{Max} > I_{Compliance}$ are caused by the current overshoot during forming. ................................. 32
2.15 Cumulative distribution plot of time taken to reach breakdown for various CVF voltage conditions. ........................................ 33

2.16 Cumulative distribution plot of HRS resistance (at -0.1V) for devices formed under different CVF conditions. ........................................ 35

2.17 HRS median and standard deviation for the devices in the figure formed at 1.5V and 1.35V CVF. ........................................ 36

2.18 Cumulative distribution plot of LRS resistance (at 0.1V) for devices formed under different CVF conditions. ........................................ 37

2.19 Time taken to reach breakdown at 1.65V CVF and corresponding $I_{Max}$ during 1$^{st}$ reset. ........................................ 38

2.20 MP TAT simulation of time taken to reach breakdown at CVF and various temperatures plotted against the corresponding current transient rate during the final breakdown phase of forming. ........................................ 39

2.21 Cumulative distribution of $I_{Max}$ for devices formed under the same CVF conditions at room temperature and 125$^\circ$C under the 100$\mu$A compliance. ........................................ 40

2.22 Cumulative distribution of $I_{Max}$ for devices formed under the low voltage CVF conditions at room temperature and 125$^\circ$C under the 100$\mu$A compliance. ........................................ 41
2.23 A comparison of the resultant median HRS for devices formed under lowe voltage CVF at high temperature compared to room temperature. Very similar behavior is observed. 42

2.24 Sketch of typical overshoot current pulse during an uncontrolled forming event. 43

2.25 Estimates of the oxygen diffusion time at various fields applied across the RRAM. Oxygen ions diffusion probability expression: lattice vibration frequency $\nu = 10^{12} s^{-1}$, jump distance $\lambda = 2.5 \times 10^{-10} m$, $Q$ is the charge of the diffusion specie ($O^-$), substitutional $O^-$ diffusion activation energy $E_A = 0.3 eV$, and $F$ is constant electric field. The characteristic diffusion time has been calculated from Monte Carlo diffusion model implementing the above formula. In this estimate, the length of diffusion needed to form a filament is assumed to be $l_D = 5 nm$. Inset: Schematic illustrating filament formation by oxygen out-diffusion with characteristic length $l_D$. 44

3.1 Sketch of a 50x50nm$^2$ crossbar RRAM cell with top electrode (TE) and bottom electrode (BE). 48
3.2 STEM Measurement: (top) Dark Field image showing a conical filament with an estimated top diameter of 5.6nm and bottom diameter of 2.5nm, (bottom) EELS elemental maps indicating a reduction of the HfO$_2$ content (left) and increase in the Hf content (right) in the filament region. [1]

3.3 Characteristic butterfly curve for the device under test (DUT) obtained using standard sweep set and reset. The plot displays the result of an initial 10 cycles of set and reset with an LRS to HRS ratio of about 10X with read at 0.1V.

3.4 Measurement setup for a) pulsed reset, and b) standard DC sweep set.

3.5 Example of a reset pulse being applied on the BE of the device in an LRS state. The applied reset pulse is followed by a ‘set’ voltage sweep from 0V to 1V and back to 0V, where the resulting HRS current is read at 0.1V to assess the effect of the reset pulse.

3.6 Filament health: monitoring stability of the HRS and LRS values throughout the characterization process.

3.7 Pulsed reset characterization flow diagram.

3.8 Complete picture of the pulsed characterization performed on a single device. Voltage-time trade-off and ‘deep reset’ phenomenon can be seen.
3.9 Impact of pulse time on HRS at fixed pulse voltages. This plot represents vertical slices through the pulsed characterization data presented in Figure 3.8.

3.10 Pulse length and voltage needed to achieve a 10X HRS/LRS ratio, depicting the voltage-time trade-off.

3.11 'Deep Reset' Phenomenon: with shorter pulses the device can withstand higher pulse voltages. The data shown in this plot was collected for two devices per pulse width condition using the pulse characterization methodology introduced in this chapter. The darker and lighter shades of each color represent the HRS and LRS, respectively. It can be seen that shorter pulses can achieve deeper resets otherwise not possible with longer pulses.

4.1 A schematic of the switching kinetics. During the forming process (left) ionized oxygen atoms out-diffuse and eventually get stored in the interstitial positions around the filament region. During the reset process (right), the O-ions diffuse following the electric field and density gradient and may re-oxidize a portion of the filament rupturing it into the high resistance state (HRS).

4.2 Temperature assisted out-diffusion of oxygen ion post forming.

4.3 Energy diagram schematic for oxygen ion diffusion in HfO₂.
4.4 Schematic of the substitutional diffusion of the oxygen ion and corresponding energy profile for the diffusion process (described by Eq. (1)). The barrier for the oxygen hopping into the vacant sites in the filament is negligible due to significant energy gain associated with the metal oxidation.

4.5 Relation of filament diameter to LRS resistance assuming bulk Hf conductivity and cylindrical filament.

4.6 Time dependence of electrical conductivity of filament ‘hafnium-like’ material used for simulation.

4.7 Temperature profile of device in LRS state right before reset starts.

4.8 Sketch of tunneling barrier implementation in COMSOL Multi-physics simulator.

4.9 Example of the electric field simulations for a filament of 5nm top and 2.5nm bottom diameter with a reset voltage of -1V applied. Poisson’s equation is solved to obtain the electric field shown.

4.10 Potential barrier between two metal electrodes.

4.11 Initial 3D ion distribution post-forming, assuming uniform distribution.

4.12 Time Dependent Monte Carlo reset simulation flow.
4.13 Two snapshots of the reset simulation progression are displayed. 1) the initial condition for O-ion distribution obtained from a random distribution of $O^{2-}$ ion simulating out-diffusion during forming 2) the final filament rupture moment, also notice the filament narrowing due to progressing of oxidation along its sidewalls.

4.14 Simulated current, temperature, and electric field simulated as a function of pulse length.

4.15 Simulated reset maximum e-field in filament area at various constant applied reset voltages plotted as a function of time.

4.16 Simulated reset maximum temperature in filament area at various constant applied reset voltages plotted as a function of time.

4.17 Simulated reset current at various constant applied reset voltages plotted as a function of time.

4.18 Three different interstitial oxygen concentration levels prior to reset. Blue region is the filament and the red dots are interstitial oxygen.

4.19 Simulated read current in filament area at various initial oxygen ion densities plotted as a function of time.

4.20 Variability simulation of reset process for 7 different cases with identical initial conditions.
4.21 Experimental data of reset resistance values as a function of applied reset pulse number. The operating range can be divided into three distinct regions as shown [2].
Chapter 1

Introduction

1.1 Evolution of Non Volatile Semiconductor Memory Technology

“The principal applications of any sufficiently new and innovative technology always have been, and will continue to be, applications created by that technology.”
— Hebert Kroemer (2000 Nobel Physics Laureate)

The invention of the fundamental floating-gate Nonvolatile Semiconductor Memory (NVSM) cell in 1967 by Kahng and Sze [3] essentially revolutionized information storage and marked the start of an era with many new applications to be created. In addition to its non-volatility, a key advantage of this invention was the creation of a single functional device rather than a circuit block (such as DRAM and SRAM) as
the fundamental unit cell in memory systems, thus reducing the silicon footprint of bit storage.

The floating-gate invention initially manifested itself in the form of the erasable programmable read-only memory (EPROM), which could only be erased by exposure to ultra-violet light, and also in the form of electrically erasable programmable read-only memory (EEPROM). In 1984 Masuoka first proposed the addition of a special erase gate to an EEPROM array so that “the contents of all memory cells are simultaneously erased by using field emission of electrons from a floating gate to an erase gate in a flash” [4] thus alleviating the issue of slower erase in EEPROM memories.

Over time, the flash floating gate memory has evolved quite a bit through various architectural advancements such as NAND introduced in 1987 [5–7] – enabling higher density, as well as through scaling – with a pace exceeding even Moore’s law for transistors [8]. However, at the moment, the aggressive scaling of NAND flash in the past 30 years has pushed the device to its physical scaling limits. Anticipation of this moment, had given rise to significant efforts on the emerging memory alternatives to flash as a flash-replacement solution.

In 2013, Samsung announced the mass-production of the first 3D vertical NAND flash memory, which they term V-NAND [9]. This innovation in flash NAND was the key to extending the life of this technology and allowing the industry to continue with the momentum of years of experience and expertise in NAND flash development. Emerging memory technologies now have to offer much more than simply enabling
further device scaling in order to be worthy of pursuit.

Over the years, a number of emerging NVM technologies have proven to be promising for future memory/storage solutions. The most promising of these NVM’s are phase-change random access memory (PCRAM) [10], spin-transfer-torque random access memory (STTRAM) [11, 12], conductive-bridge random access memory (CBRAM) [13, 14], and resistive random access memory (RRAM) [15, 16]. Device characteristics of these emerging memories have proven to be significantly better than flash giving rise to hopes of finding a “universal memory” that can function as flash replacement while also taking the place of SRAM and DRAM. The ultimate solution would achieve the switching speeds of SRAM, with the endurance of DRAM and cost/storage density of NAND flash.

The ideal characteristics for this “universal memory” would include fast programming speeds on the order of nanoseconds (<ns), low operating voltage (<1 V), low energy consumption (∼fJ/bit), long retention time (>10 years), long read/write endurance (> $10^{16}$ cycles), and excellent scalability (<10 nm) as well as 3D integratability. Satisfying the entire ideal characteristics in a single “universal memory” is extremely challenging [17], yet the aforementioned emerging memory technologies aim to at least achieve part of these ideal characteristics.

These emerging memory technologies have some features in common: they are two-terminal non-volatile devices that differentiate their states by switching between a high resistance state (HRS) and a low resistance state (LRS). The transition between
the two states is triggered by electrical inputs, yet the physical mechanism for switching is quite different for different memory: STTRAM relies on difference in resistance between the parallel configuration (LRS) and anti-parallel configuration (HRS) of two ferromagnetic layers separated by a thin tunneling insulator layer; PCRAM relies on chalcogenide material to switch between the crystalline phase (LRS) and the amorphous phase (HRS); and CBRAM/RRAM rely on the formation (LRS) and rupture (HRS) of a conductive filament in the oxide between two electrodes. This different underlying physics results in different device characteristics among emerging memory technologies.

1.2 Storage Class Memory and the Data-Centric Era

We have entered an era of “Big Data” in which the amounts of data being generated and stored is rapidly exceeding exascale [18]. Information collected by various sensors, user data, personal information, and internet all have led to this rapid increase of stored information. In fact, not only has the amount of data increased drastically, but the nature of this data has changed as well. Computation is now required to revolve around this new data and traditional algorithms of predicting more frequently accessible data are now proving less useful.

There is concern of a “memory wall” which describes the implications of the processor/memory performance gap that has grown steadily over the last several
decades. If memory latency and bandwidth become insufficient to provide processors with enough instructions and data to continue computation, processors will effectively always be stalled waiting on memory [19, 20]. Fundamentally, the von Neumann bottleneck caused by the limited bandwidth between the processor and memory in the current architecture, has become a real challenge in computing systems with this large amount of data. Unfortunately, Moore’s law scaling does not take care of this problem automatically.

Traditionally, “memory” and “storage” have had two distinct roles with ”memory” being a more temporary and low latency solution for active data needed by the processor, while “storage” being a cheaper and non-volatile solution for information storage. This has lead to the pyramid-like memory hierarchy consisting of SRAM, DRAM, hard disk, and more recently NAND flash (Figure 1.1). Going up the pyramid we have lower latency, lower capacity, more expensive memories such as SRAM and DRAM, while lower in the pyramid, we find flash and hard disk with larger capacities yet higher latency.

In addition to cost, the tip of the pyramid (“memory”) suffers from the memories being volatile and requiring refreshes to prevent the data from being lost and this leads to more power consumption. The base of the pyramid (“storage”) suffers from latency and limits computation. Interestingly enough, a convenient gap exists between this “memory” and “storage” regions in the memory hierarchy that can be filled with a memory technology without necessarily disturbing the status-quo much or replacing any incumbent memory technology. The emerging memory technologies demonstrate latency in this range and aim to fill this space in the hierarchy for a start.
Adding these levels in the memory hierarchy creates the potential to revolutionize today’s computation architecture and how processors and memory communicate. This level in the hierarchy is named “storage class memory” (SCM) in the memory field and is so called because this new memory level has “memory”-like high performance and robustness while being low cost and non-volatile like conventional “storage” [21, 22]. Geoffrey Burr suggests that the SCM can, in principle, give rise to two distinct levels in the memory pyramid essentially differentiated from each other by access time.

On the first level closer to the bottom of the pyramid a storage type (S-type) SCM would serve as a high-performance solid-state drive. This S-type SCM (S-SCM)
would be accessed by the system I/O controller similar to an HDD. The S-SCM would require a retention at least as long as flash, allowing for this SCM module to be stored offline while offering random access capability, a feature that NAND Flash devices cannot provide. These features of S-SCM would lead to improved performance and simpler systems. [23]

The second new level in the memory and storage hierarchy closer to the top of the pyramid would be a memory type SCM (M-SCM), offering a read/write latency of less than $\sim 200$ ns. This would allow the M-SCM to remain synchronous with the memory system, allowing direct connection from the memory controller and bypassing inefficiencies of access through the I/O controller. This M-SCM would play the role of increasing the memory capacity of DRAM to provide the same overall system performance as a DRAM-only system, while having moderate retention, lower power/GB, and lower cost/GB as compared to DRAM. [23]

Thus it is seen that the development of an electrically accessible nonvolatile memory with high speed, high density, and high endurance, referred to as “Storage Class Memory” (SCM) would initiate a revolution in computer architecture.

1.3 RRAM: The Resistive Switching Phenomenon in Oxides

The electrically triggered resistance change phenomenon has been observed in various metal-oxide based systems such as HfO$_2$, ZrO$_2$, Ta$_2$O$_5$, and TiO$_2$, to name a
few. Among the various mechanisms of resistance change, the filament-based system has received considerable attention due to the demonstrated nanosecond, low power (<pJ) switching with high (∼trillion cycle) endurance and retention of more than 10 years at 200C, using fab friendly, simple binary oxides and metal electrodes. [14, 24–30]

The filament-based resistance switching mechanism involves repeatable formation and rupture of conductive paths or filaments in the dielectric metal-oxide layer and has the unique attribute of area independent resistance while a filament is present. This area independence indicates an ultimate scaling advantage that is only limited by the active filament size, which can be on the order on a nm in size. The exact mechanism in filament-based switching depends highly on the materials used in the fabrication of the memory cell and can thus involve various types of conduction mechanism.

Among the various metal-oxide materials, the HfO$_2$ has recently become one of the stronger candidates for RRAM dielectric material. HfO$_2$ is a material that has now become mainstream in advanced transistor gate stack applications and thus when used for metal-oxide filament-based RRAM comes with many advantages [31, 32]. In addition to having the proven ‘fab-friendly’ track record, extensive research has been performed on HfO$_2$ from about 1997 to 2007 developing gate-stacks for advanced logic using metal-gate High-K toward manufacturability and productization. As such, detailed understanding of HfO$_2$ conductivity, and breakdown mechanisms has been obtained over these years which can be leveraged for RRAM research and development [33–36]. Moreover, Hf has one of the stronger oxygen affinities among transition metals and is relatively stable thermodynamically while being compatible with most
of the commonly used fab-friendly electrodes such as TiN, TaN, W [37–41].

Although great progress has been made in materials research, defect engineering, scaling and understanding general device operation, details of the precise kinetics of device operation and the microscopic mechanisms involved in switching operations in the HfO$_2$-based RRAM still remain unresolved, thus hindering further progress. This dissertation focuses on the HfO$_2$-based RRAM system, discussing optimization of the device operating conditions via the developed microscopic, atomistic models for possible mechanisms involved repeatable resistance switching observed. It is important to note that although some of the processes contributing to the resistive switching may be active in other metal oxide RRAM systems, due to material-specific nature of RRAM characteristics (such as relative oxygen affinities, valence states, and atom diffusivities), any conclusions or comparisons to other material systems must be made with caution [42].

1.4 Operational Characteristics of HfO$_2$-Based RRAM

The most simple RRAM memory cell consists of a metal-insulator-metal (MIM) structure for which an applied voltage across the cell is used to electrically stimulate a change of resistance. This configuration is referred to as the 1-resistor (1R) configuration (Figure 1.2).

The typical resistance change behavior for the bi-polar operation of a filament-based RRAM cell, such as HfO$_2$, is depicted in Figure 1.3. Within this switching
operation, an ohmic low resistance ‘ON’ State (LRS), or a non-ohmic High Resistance ‘OFF’ state (HRS), is related to the formation of a conductive filament (CF) or its rupture, respectively. The forming and rupture process for the CF is a random process related to the material properties associated with the dielectric material and its interaction with the metal electrodes. To describe this type of resistive switching, we must start with a microscopic description of the CF formation for which precise details that enable complete memory operations have only recently been proposed.

The CF formation process in HfO\(_2\) in essence can be described in terms of a controlled dielectric breakdown (BD) of HfO\(_2\) in which an abrupt formation of a localized region between the electrodes, within which the dielectric composition becomes more oxygen-deficient, renders this region conductive. This abrupt conductance change during the CF formation is observed during the dc voltage sweep of the RRAM dielectric ‘fresh state’ (Figure 1.3).

Once a CF is created, the ‘reset’ and ‘set’ operations rupture and re-create portions
CHAPTER 1. INTRODUCTION

Figure 1.3: Complete switching curve of a typical HfO$_2$-based RRAM device under forming, reset and set voltage sweeps.

of the filament to produce the HRS and LRS states (Figure 1.4). Figure 1.3 also indicates the dc ‘reset’ and ‘set’ operation for filament-based RRAM, as described below:

1. **Forming**: The dielectric breakdown event leading to a CF; the voltage of the abrupt breakdown during the forming voltage sweep is termed $V_F$;

2. **Reset operation**: The process in which the device switches from the LRS to the HRS state due to rupture of the CF; and

3. **Set operation**: The process where the device changes from the HRS back to the LRS state due to re-formation of the CF path.
CHAPTER 1. INTRODUCTION

Figure 1.4: Sketch of the forming and reset models in which oxygen ions are generated during forming, and interstitial oxygen ions oxidize the tip of the filament during reset.

1.5 Thesis Overview

Resistance change memory (RRAM) based on Transition Metal Oxides (TMOs), whose operation is based on the change in resistivity due to the formation of a conductive filament in the oxide material, has attracted attention in recent years because of its potential for high density, high speed, and good retention as a novel nonvolatile memory. However, achieving low power operation and high device-to-device uniformity in the cell resistance states are the major challenges for practical applications of RRAM technology. While certain progress has been made in understanding the switching mechanism of TMO memory devices, lack of precise control of the filament formation, perceived to be a random process, introduces variability into the switching characteristics of this class of devices, hindering further progress. In this thesis we address these major issues in HfO$_2$-based RRAM devices.

Chapter 1 gives a brief history of the nonvolatile semiconductor memory evolution over time starting from the invention of the floating-gate memory cell in 1967. A
discussion of Storage Class Memory and its potential revolutionary impact on computer architecture is described for the data-centric era introducing emerging memory solutions that may bring about this new level in the memory hierarchy.

In chapter 2, constant voltage forming (CVF) is proposed, shown to increase the resistances of the low resistance and high resistance states while reducing their variability. By forcing the forming in all devices to occur at the same predefined voltage, the CVF method is demonstrated to eliminate a major cause of the device-to-device variation associated with the randomness of the forming voltage values. Moreover, both experiments and simulations show that CVF at lower voltages suppresses the parasitic overshoot current, resulting in a more controlled and smaller filament cross-section and lower operation currents.

Electrical and physical characterization of the HfO$_2$ RRAM devices is demonstrated in chapter 3, presenting physical evidence for a filament in HfO$_2$ RRAM using scanning transmission electron microscopy. Moreover, a ‘deep reset’ phenomenon is observed electrically for ultra-short pulses at larger voltages otherwise not observable using longer pulse times. The interplay of breakdown and recombination forces for hafnium and oxygen ions is suggested to explain this observed phenomenon.

Chapter 4 presents a detailed model for the reset process kinetics in HfO$_2$-based RRAM describing the transition between low and high resistance states at the atomic level. Based on the filament characteristics as observed by TEM, the kinetics of the reset operation is then simulated using our developed Kinetic Monte Carlo (KMC) method incorporating ab-initio calculated microscopic characteristics of the oxygen
ions in hafnia. Temperature and field driven oxygen diffusion in the oxide surrounding the filament is shown to provide the needed supply of oxygen to re-oxidize the tip of the filament and switch the device back to the High Resistance State (HRS). Using this developed model and KMC simulator, the dependence of reset process on surrounding interstitial oxygen ion concentration, reset pulse width, and reset pulse height is studied and the variability of the resulting HRS state is simulated.

Chapter 5 summarizes the results and contributions of this thesis, and ends the chapter with an overview of future work proposed.
Chapter 2

The Forming Process: Controlling Uniformity of RRAM Characteristics

While certain progress has been made in understanding the switching mechanism of TMO memory devices [14, 16, 43, 44], the random nature of the atomic structure of RRAM cells introduces variation in device operation which needs to be controlled for RRAM to become a practical solution. Primarily, lack of precise control of the filament formation, perceived to be a random process, introduces variability into the switching characteristics of this class of devices, hindering further progress. The forming step, as will be demonstrated in this chapter, is a key initiation step that configures the RRAM cell for subsequent switching from LRS to HRS and back to LRS.
In this chapter we delve into the details of the forming process and demonstrate a forming methodology, which addresses the issue of uncontrolled forming and variability in the device characteristics. We start by looking into the sources of conductive filament variability caused by the filament forming process, and then take advantage of parasitic capacitance inherent in our circuit to experimentally characterize the forming transient time. As we are dealing with an Metal-Insulator-Metal (MIM) structure, we systematically reduce our applied forming voltage \( V_F \) to perform forming and obtain filaments at various \( V_F \) values albeit at the cost of time to forming.

As a result, the forming voltage impact on the filament creation process is studied. The proposed methodology performs the forming operation under the constant voltage stress (CVS) condition at lower voltages rather than using the conventional fast voltage ramp method. This approach is shown to lower the reset current, increase the resistivity of the low and high resistance states (LRS, HRS), and improve device-to-device uniformity in HfO\(_2\)-based RRAM devices. Multi phonon trap assisted tunneling forming simulations are employed to simulate the observed trends from our experiments.

### 2.1 Sources of Conductive Filament Variability

Hafnium oxide based RRAM cells having a stacked structure of \( TiN/Zr/HfO_2/TiN \) RRAM were characterized using the 1T1R setup in which an external transistor is connected in series with the cell. The purpose of the transistor is to control the compliance current during forming and subsequent set operations. Alternatively limiting
the maximum current supplied to the circuit has also been done with the Semiconductor Parametric Analyzer (SPA) by employing the current compliance limit circuitry, however the latency of the SPA in limiting the current is much larger than the forming event and thus this method has its limits.

The series transistor can alleviate this problem by almost instantaneously shifting the voltage drop across the resistor to drop across the transistor and thus limit the maximum current during forming as desired. The forming event in which the resistance drops from the "fresh state" resistance to "low resistance state" occurs in a very short time interval and is not straightforward to measure especially without impacting the forming process itself.

The maximum current during the $1^{st}$ reset after forming ($I_{Max}$) is found to be a reliable figure of merit for the properties of the formed filament (e.g. size and resistivity); as seen in Figure 2.1, devices in which the $I_{Max}$ values are above the compliance current limit exhibit lower LRS resistance values (leading to greater power consumption during reset) and a higher variation in the HRS resistance during subsequent device cycling (not shown). Since a greater $I_{Max}$ likely reflects a larger effective filament cross-section (with less resistivity), a forming process resulting in lower $I_{Max}$ values is highly desirable for achieving low power operations.

The distribution of forming voltage values across a set of identical devices obtained by forming the devices by the commonly used voltage sweep forming method is depicted in Figure 2.2. Further, the maximum current during the first reset after forming is measured, and a correlation between higher forming voltages ($V_F$) and
CHAPTER 2. CONTROLLING RRAM CHARACTERISTICS

Figure 2.1: Examples of devices with (left) overshoot during forming as manifested by the $I_{Max}$ value being above the compliance limit, and (right) without overshoot. Maintaining a compliance limit during forming results in higher LRS and HRS resistances.

Higher $I_{Max}$ is observed, corroborating previous reports [45]. Figure 2.3 illustrates this correlation and thus the importance of reducing both the magnitude and variability of $V_F$ to improve device-to-device uniformity [46].

As depicted in Figure 2.3, the device-to-device variation of $I_{Max}$ values can be decomposed into two major contributions:

1. “Extrinsic” variation, due to the random distribution of the $V_F$ values produced by the voltage ramp forming process.

2. “Intrinsic” variation, due to additional factors determining the conduction characteristics of the filament created at a given forming voltage $V_F$.

We, therefore, propose a constant voltage forming (CVF) process that eliminates (by definition) the first contribution to $I_{Max}$. As shown below, CVF performed at low voltages also reduces the second contribution.
CHAPTER 2. CONTROLLING RRAM CHARACTERISTICS

Figure 2.2: Usual distribution of forming voltages across different devices during a typical forming sweep.

2.2 Caused by Compliance Current Overshoot

Due to parasitic capacitance ($C_P$) in the RRAM external 1T1R circuit, the maximum current through the RRAM device during forming may surpass the current compliance limit set by the transistor (Figure 2.4) as a result of the AC path through the $C_P$. This parasitic $I_{C_P}$ current becomes more dominant (relative to the current compliance) with lower compliance limits needed to obtain higher LRS and HRS resistances [47]. Increasing the performance of the memory cell by achieving faster switching speeds, for set (HRS to LRS) for example, will amplify the impact of this parasitic current $I_{C_P}$ as well and thus it becomes imperative to understand and control this effect.
Figure 2.3: Dependence of the maximum of the 1st reset current ($I_{Max}$) on the forming voltage, $V_F$.

In our experimental setup with the circuit shown in Figure 2.4, the value of $C_P = \sim 1nF$ (which is relatively high, thus magnifying its effect) was extracted by applying a constant current at node $V_1$ while disconnecting the DC path to ground and measuring the slope of the voltage increase. The transient (overshoot) current, $I_{C_p}$, induced by this parasitic capacitance is proportional to $\frac{dV}{dt}$, the rate of the voltage decrease across the cell during the dielectric breakdown instance (the final, current-runaway forming phase) as given by $I_{C_p} = C_P \frac{dV}{dt}$. The voltage at node $V_1$ in Figure 2.4 was measured using an oscilloscope; the derivative $\frac{dV}{dt}$ was then calculated to obtain the transient current values.
Figure 2.4: Schematic of the measurement setup. The transient current $I_{C_p} = C_p \frac{dV_1}{dt}$ (charging of the parasitic capacitance $C_p$) is estimated using the $\frac{dV_1}{dt}$ values as measured by the oscilloscope (at node $V_1$).

In this experiment, four devices were tested under identical CVF conditions of 1.7V with a 100\(\mu\)A compliance limit. The peak value of $I_{C_p}$ is observed to match well to the $I_{Max}$ (Figure 2.5) indicating that the pre-set compliance limit determined by the transistor in series with the RRAM cell is not effective in controlling the current overshoot. The current overshoot effectively increases the current compliance limit, which has been shown to control the filament cross-section [16, 48], thus introducing an undesired, uncontrolled random component in the forming process.
Figure 2.5: (a) Examples of the overshoot currents and the corresponding $I_{Max}$ values - $I_{Max}$ (peak current) is measured in the first reset after forming- (the “intrinsic” variation illustrated in Fig 2.3) for 4 devices during the forming event under the identical CVF condition of 1.7V with the 100$\mu$A compliance. (b) The 1$^{st}$ reset (negative voltage) and set (positive voltage) of the corresponding devices.

### 2.2.1 Modeling the Impact of $C_P$ on Overshoot Current

For simplicity, the decrease in RRAM resistance during the dielectric breakdown (final) phase of the forming process is approximated by an exponential time dependency as shown in Figure 2.6 (note that the actual process follows a more complex time dependency, as the simulations in the next section will show),

$$R(t) = (R_{Max} - R_{LRS}) \times e^{(-\frac{t}{t_r})} + R_{LRS}$$

where $R_{Max}$ is the resistance of the RRAM device in the initial (fresh) state, $R_{LRS}$ is its resistance after forming, and $t_r$ is the breakdown transient time. The current
through the RRAM device can then be obtained through,

\[ I_{RRAM}(t) = \frac{V_F - V_1(t)}{R(t)} \]

where \( V_F \) is the voltage applied during forming, and \( V_1 \) is the voltage at node \( V_1 \) as shown in Figure 2.4, which is calculated by solving,

\[ C_F \frac{dV_1}{dt} + V_1(t) \left( \frac{1}{R(t)} + \frac{1}{R_L} \right) = \frac{V_F}{R(t)} \]

\( R_L \) is the resistance seen at the drain of the transistor, which includes the MOSFET channel resistance and the 50Ω resistance connected at the source.

![Simulated final breakdown phase of forming approximated by an exponential time dependence.](image)

Figure 2.6: Simulated final breakdown phase of forming approximated by an exponential time dependence.

The fitting of the measured overshoot current through the RRAM yields a characteristic time constant of 200ns for the resistance change (Figure 2.7) for this particular device. Within this description, the impact of shorter RRAM forming transients on
the increase in overshoot current is simulated (Figure 2.8). The results clearly show
the existence of a direct correlation between $t_r$ and the overshoot magnitude, which
increases significantly as the final breakdown forming phase becomes faster.

![Current through RRAM during Forming Process](image)

Figure 2.7: An example of the modeling (broken line) of the overshoot current (device 2 in Fig. 2.5) through RRAM. The parasitic capacitance is estimated to be around 1nF (using an external transistor in the 1T1R setup).

While the forming transient time, $t_r$, depends on the physics of the forming process
and can be controlled to a certain degree by the forming conditions, the value of $C_P$
can be reduced through proper device integration. Figure 2.9 shows how reducing $C_P$
can help reduce the overshoot magnitude; however, faster resistance transients and a
more stringent requirement for lower compliance limits may still lead to overshoot.
2.3 Controlling the Filament Formation Process

The characteristics of the RRAM cell stem from the initial filament created during the filament formation process. This filament creation process in turn is a very delicate process that depends on the device material, atomic structure, defects as well as the kinetics of how the “controlled dielectric break-down” is performed during the filament formation process. Details of the kinetics of the switching are discussed in future chapters. Here we focus on the effects of external stimulants on the breakdown instant, which can be utilized to control the forming process, namely forming voltage ($V_F$) and ambient temperature during forming.
Chaprer 2. Controlling RRAM Characteristics

2.3.1 Simulating Impact of Forming Conditions on Breakdown Transient Rate

To understand how forming voltage and temperature affect the intrinsic (in the MIM device with no parasitics) BD transient rate \( \left( \frac{dI}{dt} \sim -\frac{dR}{dt} \right) \), simulations of the forming process were performed (Figure 2.10) using the RRAM statistical forming model [48, 49]. The model assumes multi-phonon trap-assisted tunneling via positively charged oxygen vacancies as the primary conduction mechanism through the HfO\(_x\) dielectric [16, 50], while accounting for the structural lattice relaxation associated with the electron capture and emission events at the defects.

The field- and temperature- dependent defect generation probability during Constant Voltage Forming (CVF) is calculated from the power dissipation at individual
CHAPTER 2. CONTROLLING RRAM CHARACTERISTICS

trap sites and the corresponding local temperature increase according to:

\[ G(x, y, z) = G_0 \exp\left( -\frac{E_A - \beta E_{ox}(x, y, z)}{kT(x, y, z)} \right) \] (2.1)

where \( G \) is the probability of breakdown of any particular HfO\(_2\) bond at the location \((x,y,z)\), \( G_0 \) is the characteristic vibration frequency of the bond, \( E_A \) is the breakdown energy barrier, \( E_{ox}(x, y, z) \) is the applied electric field, \( \beta \) is the constant that accounts for the polarization of the chemical bonds by the electric field, \( k \) is Boltzmann’s constant, and \( T(x, y, z) \) is the local temperature.

The leakage current through the dielectric is associated with injected electrons hopping through the vacancies. The process of the electron capture and emission (trapping/detrapping) at a vacancy site during this hopping occurs with a lattice relaxation of the surrounding atoms resulting in an energy dissipation that effects the local temperature surrounding the vacancy. This local temperature change is captured in the simulation and impacts the generation probability of new defects as accounted for in equation 2.1.

The model also accounts for the transition of the local conduction mechanism from the TAT to drift, which occurs when the defect density in the oxide exceeds a critical density resulting in the formation of a conductive sub-band [48]. The charge transport as well as the subsequent formation of a conductive filament is assumed to occur at the grain boundaries.

Twenty devices with randomly selected initial trap distributions along the grain
boundaries were simulated at various stress voltages and ambient temperatures under the Constant Voltage Stress (CVS) conditions. Figure 2.10 shows the produced forming curves from these simulations under the 225°C temperature condition and 2V CVS. The intrinsic breakdown transient rate \( \frac{dI}{dt} \) is defined as the time it takes for the current to raise from the initial (arbitrary) level of 1\( \mu \)A to the compliance current or in this case the final runaway level. The simulations show that intrinsic BD transient rates are higher when the CVF voltage (Figure 2.11) and temperature (Figure 2.12) are higher.

Figure 2.10: An example of the TDDB simulations using the forming model: CVF at 2V, T=225°C, 20 devices with randomly selected initial trap distributions. For each device (each line), the current transient rate \( \frac{dI}{dt} \) is defined by the time required for the current to grow from the initial (arbitrary) level of 1\( \mu \)A to the final BD (compliance) current.

At higher \( V_F \), the defect generation rate governed by equation 2.1 is higher leading
to shorter BD transient times. In addition, at higher temperatures, a given current magnitude can be reached with fewer traps in the filament, Figure 2.12, due to a higher rate of electron transfer through the traps [50]. Therefore, any newly generated trap results in a greater current increase (than at lower temperatures) resulting in a higher $\frac{df}{dt}$. Thus, higher values of both externally controlled parameters, $V_F$ and $T$, tend to shorten the transient time, leading to a greater parasitic $I_{CP}$ current (as seen in Figure 2.8) and, subsequently, a larger filament.
Figure 2.12: Simulated $\frac{dt}{dt}$ dependence on ambient temperature at 2V CVS, as well as number of traps needed in the filament to reach the forming compliance; a weaker dependence than on the forming voltage is observed.

2.3.2 Measurement of Forming Conditions Impact on Device Characteristics

To experimentally verify the impact of the forming voltage and temperature on the BD transient rate, and thereby the filament formation, forming with various CVF conditions was performed on groups of 20 devices per CVF condition. The initial leakage current at each set of devices was obtained and the different groups of fresh devices were compared to ensure the initial samples are relatively similar (Figure 2.13). The constant voltage was applied on the devices using an Agilent 4156C Precision Semiconductor Parameter Analyzer. The device was thus stressed
using the external 1T1R configuration for as long as needed until the breakdown would occur, at which point the transistor would limit the maximum current through the device. The time it took from the initial stress until the breakdown point would occur, or the time-to-forming, was recorded as well to study any possible correlation with device characteristics.

As expected, and in accordance with the time-dependent dielectric breakdown (TDDB) in MOSFETs, the MIM RRAM device takes a longer time to form under lower voltage conditions, yet despite being formed with various stress conditions, subsequent set and reset is observable. The maximum current during the 1st reset after forming was measured and impact of the CVF condition on the formed filament was studied. As seen in Figure 2.14, the CVF at lower voltages suppresses $I_{Max}$.

Figure 2.13: Initial Leakage of randomly selected group of fresh devices before any forming.
significantly. This is done at the cost of a longer time-to-form (i.e. the overall time-to-form the device, which should not be confused with the breakdown transient time controlling the overshoot). The average time-to-form for the different CVF conditions ranges from tens of milliseconds for the 1.65V condition to about 750s for the 1.35V condition and is depicted in Figure 2.15.

![Cumulative distribution of $I_{Max}$ for devices formed by different CVFs under 100$\mu$A compliance. The values of $I_{Max} > I_{Compliance}$ are caused by the current overshoot during forming.](image)

Figure 2.14: Cumulative distribution of $I_{Max}$ for devices formed by different CVFs under 100$\mu$A compliance. The values of $I_{Max} > I_{Compliance}$ are caused by the current overshoot during forming.

The elimination or significant reduction of the overshoot current during the forming step has at least two advantages. On the one hand, the maximum current through the device is reduced and a smaller filament is formed therefore helping to achieve low power forming and potentially lower power operation. And on the other hand, part of the *intrinsic* variability in the filament formation process is controlled thus
Figure 2.15: Cumulative distribution plot of time taken to reach breakdown for various CVF voltage conditions.

As it is observed in Figure 2.14, the reduction in CVS voltage and thus reduction in $I_{\text{Max}}$ level also produces more uniformity in the $I_{\text{Max}}$ levels.

The next question that remains to be answered is the impact of this CVS forming method on the subsequent operation of the device. To address this issue, each device is cycled 10 times with a typical sweep set and reset. The current compliance limit during the 'set' is kept at the 100μA level and applied using the external 1T1R configuration as in the forming step and the 'reset' occurs with a sweep up to -1.2V. The result for each device is a butterfly curve similar to one depicted in Figure 2.1.

A few figures of merit for each of such curves are chosen to compare the forming
conditions. These figures of merit in addition to the already presented $I_{Max}$ are: a) LRS resistance read at 0.1V, and b) HRS resistance read at -0.1V. We look at the median to exclude any possible outliers as well as the standard deviation for the set of 10 cycles. In this manner, two types of variations can be studied, namely cycle-to-cycle variation, and device-to-device variation for each Constant Voltage Forming (CVF) condition.

The resistance of the HRS calculated as $\frac{0.1V}{\text{current at } -0.1V}$ is plotted for each device as the median HRS of 10 reset cycles. Figure 2.16 depicts the cumulative distribution of this median HRS for the set of devices in each CVF group. As can be observed, the spread of the median reduces and a tighter distribution is obtained with the voltage reduction in CVF. This is attributed to the reduction in variability of the filament shape obtained due to the gained control from overshoot current suppression. It is important to note that the variability in this initial filament shape is seen to propagate to subsequent switching operations beyond the forming step as observed from the HRS, emphasizing the importance of controlling the forming operation.

The other significant finding regarding the HRS resistance is the spread of the HRS within the cycling for each device. As seen from Figure 2.17, the standard deviation of the HRS resistance per device reduces with the CVF voltage reduction. This result indicates that the filament created with the lower voltage, albeit with a longer time to form and a slower process, shows more stability towards switching operation. The reasons for this stability are not intuitive and have to be searched for in the kinetics of the forming operation and how the filament, vacancies and Oxygen ions are configured for subsequent switching. We will discuss this kinetics in much
CHAPTER 2. CONTROLLING RRAM CHARACTERISTICS

Figure 2.16: Cumulative distribution plot of HRS resistance (at -0.1V) for devices formed under different CVF conditions.

detail in Chapter 4.

Lowering the CVF voltage also helps in reducing the device operating voltage by reducing the LRS resistance during the device cycling. Figure 2.18 shows how reducing the forming voltage to 1.35V helps increase the LRS resistance after a set operation and allow for less current consumption during both set and reset operation. This is essentially a byproduct of gained control on the filament creation achieved by the reduction of the overshoot current during the forming operation. The spread of the LRS resistance is also reduced at lower CVF voltages as also seen from Figure 2.18.
As has been shown, the CVF method with lower voltage levels produces various benefits in terms of gained uniformity and lower power operations by trading the variability in forming voltage levels with variability in time-to-form duration. The variation in time-to-form itself may have a broad range and one concern is how this might affect the filaments created under the different time-to-form times. To address this, we look at the $I_{Max}$ parameter as a function of time-to-form to see if any such correlation can be established. Figure 2.19 is a scatter plot of the time-to-form data for the CVF = 1.6V case plotted against the maximum current during first reset after forming ($I_{Max}$). As is evident, no significant correlation is observed experimentally.

In fact, our multi-phonon trap assisted tunneling forming simulations also predict
no strong correlation between time-to-form and the rate of dielectric breakdown even at different temperatures (see Figure 2.20). Therefore, although the time to reach the ideal conditions for a breakdown (forming) to occur varies significantly from device to device, the actual forming itself (fast runaway phase) happens in a similar fashion under similar CVF conditions.

The major disadvantage of the CVF method is the increased time it takes for forming the devices when lower voltage CVF is required. CVF voltages of 1.35V can take up to 1000 seconds to form which may not be practical in real world applications. To solve this problem, forming at higher temperatures is proposed. From Equation 2.1 we know that the ambient temperature will affect the probability of defect generation.
Figure 2.19: Time taken to reach breakdown at 1.65V CVF and corresponding $I_{Max}$ during 1st reset.

and that a higher temperatures will result in faster breakdowns in the oxides.

Higher temperature forming [31] can reduce the time-to-form down to 150ms when the 1.35V CVF is performed at 125C. Although higher temperature does impact the overshoot magnitude, when comparing similar CVF voltages, a slight reduction in the forming voltage is sufficient to render insignificant this temperature impact on overshoot current, mainly due to the weaker $\frac{dT}{dt}$ dependency on temperature compared to the voltage dependency (see Figures 2.11, 2.12). Figure 2.21 compares the impact on overshoot current for two CVF conditions with similar voltages of 1.65V and different forming ambient temperatures of room temperature and 125C. Both sets of devices still display significant overshoot above the 100µA level, but the $I_{Max}$ in the
CHAPTER 2. CONTROLLING RRAM CHARACTERISTICS

Figure 2.20: MP TAT simulation of time taken to reach breakdown at CVF and various temperatures plotted against the corresponding current transient rate during the final breakdown phase of forming.

Increasing the ambient temperature during forming not only helps with reducing the time-to-forming for the devices but further allows for reduction in the forming voltage to levels not previously practical. Figure 2.22 shows the impact of reducing the forming voltage at high temperatures to the 1V level. Notice that a tighter distribution is achievable with this lower voltage. Also notable is the slight increase in the $I_{Max}$ values for the CVF 1.35V at higher temperature. A look at the figures of merit during cycling for each of these conditions does not show any significant change in operation (Figure 2.23). Thus higher temperature forming may help reduce the
CHAPTER 2. CONTROLLING RRAM CHARACTERISTICS

Figure 2.21: Cumulative distribution of $I_{Max}$ for devices formed under the same CVF conditions at room temperature and 125°C under the 100 µA compliance forming voltage without any disadvantage.

2.4 Oxygen Ion Diffusion during the Forming Process

The filament formed under the overshoot condition is expected to be less robust than one formed under the current compliance condition and may result in greater device-to-device non-uniformity. In particular, estimates of the characteristic times of oxygen diffusion in HfO$_2$, Figure 2.25, assuming the substitutional diffusion of the
oxygen(-) ion (which exhibits the lowest activation energy of 0.3eV [35]), suggest that while more O-Hf bonds must have been broken to form a region with high density of unoxidized metal atoms (a conductive filament) to support a large overshoot current through the device, the timescale of this large overshoot current might not be sufficient for the oxygen to diffuse out of the filament region (Figure 2.24). As a result, one may speculate, the filament might partially reoxidize, depending on the specifics of the temperature profile during the BD event in a given device. This would contribute to the device-to-device non-uniformity of the resistance characteristics. In agreement with the above considerations, a higher $I_{Max}$ results in higher HRS non-uniformity, as seen in Figure 2.16.
Figure 2.23: A comparison of the resultant median HRS for devices formed under low voltage CVF at high temperature compared to room temperature. Very similar behavior is observed.

Thus, the CVF method has been demonstrated to produce the desired higher resistance of the LRS and HRS while reducing their variability. This is achieved by assuring that the filament is formed under the lowest voltage practical, thus allowing for a reduction in the current compliance (limited by the leakage current through the dielectric film) during the forming process. By forcing the forming in all devices to occur at the same predefined, low voltage, the CVF method eliminates a major cause of the device-to-device variation associated with the randomness of the VF values (the source of “extrinsic” distribution of $I_{Max}$). At the same time, forming at low voltages
suppresses the overshoot current associated with parasitic capacitance, which might be present in the RRAM circuit, resulting in a smaller filament cross-section and less overshoot-related variability. It is important to note that the parasitic capacitance consideration will always be critical as a ‘set’ $I_{\text{compliance}}$ of 1$\mu$A with switching from HRS to LRS occurring in less than 1$\text{ns}$ will require a maximum $C_P$ of only $\sim 1\text{fF}$ levels (assuming set occurs at 1V)!

---

**Figure 2.24**: Sketch of typical overshoot current pulse during an uncontrolled forming event.
Figure 2.25: Estimates of the oxygen diffusion time at various fields applied across the RRAM. Oxygen ions diffusion probability expression: lattice vibration frequency $\nu = 10^{12} \, s^{-1}$, jump distance $\lambda = 2.5 \times 10^{-10} \, m$, $Q$ is the charge of the diffusion specie ($O^-$), substitutional $O^-$ diffusion activation energy $E_A = 0.3 \, eV$, and $F$ is constant electric field. The characteristic diffusion time has been calculated from Monte Carlo diffusion model implementing the above formula. In this estimate, the length of diffusion needed to form a filament is assumed to be $l_D = \sim 5 \, nm$. Inset: Schematic illustrating filament formation by oxygen out-diffusion with characteristic length $l_D$. 

$$R_{Diff} = \nu \times e^{\left(\frac{-q(E_A - Q \frac{\lambda}{2} F)}{kT}\right)}$$
Chapter 3

Physical and Electrical Characterization

Understanding the physical mechanism of the switching process in RRAM devices requires very careful and accurate characterization of the devices both physically and electrically. Since the typical operation mechanism of an RRAM device is much different from previous CMOS devices fabricated, much of the common characterization tools are not designed or optimized to capture the details of the RRAM switching operation. In particular, the following three unique features make this characterization very challenging.

Firstly, the switching between the fresh state of a device to LRS and then to HRS requires a resistance change of many orders of magnitude. As explained in the previous chapter precise control of the parasitic capacitances plays an important role
in being able to accurately control or even measure the exact current going through the devices during the resistance change. In addition, precisely limiting the maximum current through the device using the semiconductor parametric analyzer is not trivial when the current range is changing so much and also given the limited response time of the measurement system.

Secondly, and more importantly, as we have seen in the case of forming, the speed at which this resistance change occurs is much faster than hundreds of nanoseconds. This is especially true when higher switching speeds are desired, as switching speeds depend on the operating voltage as seen in the previous chapter. In fact, switching speeds of sub-nanoseconds have been demonstrated on Hafnium Oxide based systems [51]. Switching speeds on the order of faster than nanoseconds imply that we are essentially dealing with a high frequency system with frequencies higher than the GHz range. Conventional characterization of the RRAM devices with coaxial/triaxial cables and probe tips are designed with specifications of bandwidth limitations of around 150MHz assuming correct impedance matching at the device under test [52]. Thus, accurate high speed characterization of the RRAM systems which operate at these higher speeds requires special high frequency considerations in the device design phase.

Finally, due to the small dimension of the filament and random nature of the filament creation itself, physical observation of the change in the RRAM device due to state change has not been observed for most metal oxide devices including HfO$_2$. The RRAM device is very sensitive to changes in defect density, oxygen content, and as we have seen, parasitic components. Thus, in addition to identifying the precise
location of the RRAM to sample for physical characterization, the mentioned factors make it difficult to image the device without affecting its operation. However despite its challenges, in this chapter we present some physical evidence of the observation of a Hf-rich filament in our HfO$_2$ based RRAM devices.

In the following sections, we characterize the HfO$_2$ RRAM devices electrically and physically, keeping in mind the above-mentioned limitations and challenges.

### 3.1 Observation of Conical Filament in HfO$_2$ RRAM

The physical characterization of the HfO$_2$ base RRAM devices was performed in collaboration with the Institute of Microelectronics and Microsystems (IMM) in Italy [53]. Transmission Electron Microscopy (TEM) was performed using a JEOL JEM2010 equipped with Electron Energy Loss Spectroscopy (EELS) and Scanning TEM (STEM) imaging capabilities. In this study, two different techniques of STEM dark field imaging and EELS at low energy are used. STEM dark field is sensitive to the local average atomic number while EELS is strictly related to plasmon losses, determined by the local chemical composition and phase. The lateral resolution of the resulting is about 1nm, limited mainly by the STEM electron beam size.

The samples prepared for this study were from devices fabricated in the crossbar configuration with 50nm thick metal arrays producing 50x50 nm$^2$ device sizes (Figure 3.1). The device stack used was a 5nm thick HfO$_2$ switching layer with TiN top and bottom electrode and a thin Hf metal layer above the oxide film used as an oxygen
exchange layer for oxygen gettering and defect generation. The devices was ‘formed’ using the conventional DC voltage sweep at room temperature with a compliance current limit of 1mA. The devices were then tested to ensure that switching is achievable before imaging was performed.

Figure 3.1: Sketch of a 50x50nm$^2$ crossbar RRAM cell with top electrode (TE) and bottom electrode (BE).

Achieving an RRAM device capable of switching is a fairly delicate task requiring the correct oxygen content and device encapsulation from the environment. As such, and to avoid any change in the filament formed in these devices, care was taken to keep the device intact during sample preparation and so the entire device was included in the TEM specimen. Focused Ion Beam (FIB) was used to prepare the sample for TEM. The sample was prepared by cutting a vertical slice with a thickness of 60nm, just thicker than the 50nm device itself. Thus the TEM was done on the side view of the device.

The scanning TEM (STEM) dark field micrograph if the device prepared as described is shown in Figure 3.2. A bright conically shaped filament having 2.5nm and 5.6nm diameters next to the cathode and anode electrodes, respectively, is visible in
the HfO₂ oxide layer. The contrast in this image is determined by the local electronic density, and thus by the local atomic number Z per unit volume. A brighter area in this image, therefore, corresponds to a region in the insulating layer with higher average Z. The brighter region observed in the HfO₂ layer can be explained by a metallic Hf-rich filament being formed in the region.

Figure 3.2: STEM Measurement: (top) Dark Field image showing a conical filament with an estimated top diameter of 5.6nm and bottom diameter of 2.5nm, (bottom) EELS elemental maps indicating a reduction of the HfO₂ content (left) and increase in the Hf content (right) in the filament region. [1]

In an attempt to corroborate this z-contrast dark field STEM, Electron Energy Loss Spectroscopy (EELS-STEM) imaging was also been performed. The EELS spectra were collected point-by-point ( < 1nm diameter electron beam spot with the pixels of 1x1 nm²) throughout the entire image area. Electron energy loss at low energy is
mainly dominated by plasmon losses which is sensitive to the local chemical composition. Figure 3.2 shows the EELS spectra collected in the region containing either Hf, or HfO$_2$. Each spectrum can be fitted by a sum of Gaussian functions, whose peak position, width and height are characteristic of the specific material. Due to a higher scattering factor of the atoms with higher atomic number, such as Hf, the signal intensity acquired in the regions containing pure metallic Hf is correspondingly much lower [53, 54].

Figure 3.2 shows, from left to right, the maps of the normalized coefficients of HfO$_2$ and metallic Hf, as determined from the fitting of the EELS spectra in each pixel. The lack of the HfO$_2$ signal is observed around the central region of the device dielectric, with a corresponding increase of the metallic Hf signal. The data shows that the filament is of a conical shape with the estimated top and bottom radii of 5.6nm and 2.5nm, respectively. This is in good agreement with the Z-contrast dark-field STEM.

Thus the STEM analysis with two different techniques indicate the presence of a filament of the metallic Hf in the device undergone the forming process. This conclusion is in a good agreement with the model proposed in [16], where the conduction in Hf oxide is described by multi-phonon trap assisted tunneling (TAT) and a conical filament with the narrower end formed near the cathode is expected. Indeed the electron transfer in the TAT process allows more energy to be released at the traps in the dielectric region farther away from the e-injecting cathode. This result forms a basis for our modeling efforts presented in the next chapter.
3.2 Kinetics of RRAM Switching using Ultra-Short Pulse Characterization

As we have seen from chapter 2 the physical mechanism leading to the forming operation is one involving local electric field and local temperature at defect sites leading to a runaway process similar to much studied breakdown in dielectrics. Conventional characterization of RRAM devices using millisecond or longer applied voltage and current levels can serve as good methods for general device comparison or for first order understanding of device operation. However, understanding the underlying physical mechanism at the atomic level requires more precise characterization at the time scale and energy levels on the order of that needed for switching in RRAM devices. Moreover, real RRAM applications demand low power and high speed operation achieved through input pulses of short lengths, and thus understanding the device characteristics at short time scales becomes critical.

Previous work on using pulsed mode operation of RRAM devices shows promising results in terms of the potential sub-nanosecond speed [51] at which RRAM devices can be operated at, or the voltage-time trade-off that is present when the pulse energy is reduced. However, a systematic methodology with ultra-short pulse characterization accounting for circuit parasitic components, capable of capturing device operation kinetics is still lacking.

In this chapter a pulsed characterization methodology is proposed to mimic realistic device operation as well as to more accurately capture the kinetics of the switching
process. The methodology can be used for both the ‘set’ process as well as the ‘reset’ process, however, we focus on the ‘reset’ process for two reasons. Firstly, as we saw in the previous chapter, the kinetics of the ‘forming’ is very similar to breakdown in dielectric material which has been studied for decades and less unknown. The ‘set’ process follows very similar kinetics to the ‘forming’ process. Secondly, and more importantly, the breakdown nature of the ‘set’ and ‘forming’ process results in orders of magnitude of resistance decrease in very short time scales needing precise control of parasitic behavior difficult to achieve at short pulse lengths. The ‘reset’ case is more interesting and less understood, as can be noted from the many proposed models in literature [1, 42, 55–65].

Figure 3.3 shows the typical switching curve for a device ‘formed’ with a 100µA current compliance level. The positive voltages show the ‘set’ from a HRS state to LRS, while the negative voltages show the gradual process of LRS to HRS ‘reset’ occurring. We can read the state of the device at 0.1V to determine if it is in LRS or HRS, and an approximately 10X window can be seen for this particular device. This device is used to perform pulsed characterization with the methodology as proposed bellow.

The device has been ‘formed’ in an external 1T1R configuration to control overshoot current as discussed in Chapter 2, however since the ‘set’ occurs with a slow sweep with voltage steps one order of larger than a few milliseconds and since the set voltage is much smaller than the $V_F$, overshoot during set is not significant and so a 1R configuration is used to reduce parasitic effects from switch box, cabling, and transistor during the relatively high frequency ‘reset’ pulse. The measurement setup
Figure 3.3: Characteristic butterfly curve for the device under test (DUT) obtained using standard sweep set and reset. The plot displays the result of an initial 10 cycles of set and reset with an LRS to HRS ratio of about 10X with read at 0.1V.

for the reset and set is shown in Figure 3.4.

![Diagram](image)

Figure 3.4: Measurement setup for a) pulsed reset, and b) standard DC sweep set.

In the reset pulsed characterization methodology, the reset process is performed by applying a specific pulse with a particular voltage level and time duration. Figure 3.5 depicts one such reset pulse on the left of the figure. In order to assess the impact of that particular pulse, a conventional ‘set’ using a voltage sweep is performed (left
side of Figure 3.5) and the current at 0.1V on the upward sweep (before a ‘set’ event) is read and this resulting HRS is associated with the applied pulse (right side of Figure 3.5). This sweep ‘set’ also initializes the filament back to the initial LRS level. The variability in filament reset is accounted for by repeating this process 10 times per pulse condition. The median of this distribution is a better reflection of the reset pulse impact.

**Pulsed Reset**

![Graph showing pulse and set sweep](image)

Figure 3.5: Example of a reset pulse being applied on the BE of the device in an LRS state. The applied reset pulse is followed by a ‘set’ voltage sweep from 0V to 1V and back to 0V, where the resulting HRS current is read at 0.1V to assess the effect of the reset pulse.

The entire set of pulse conditions with various voltages and durations are applied on the same device to reduce any variability that might be due to device to device variation. As such, it is critical to monitor the health of the device throughout the pulsed characterization process to ensure that the LRS or HRS values are not changing over time. Figure 3.6 shows how the filament stability is monitored in between different applied pulse conditions.
Figure 3.6: Filament health: monitoring stability of the HRS and LRS values throughout the characterization process.

The flow chart for the entire reset pulsed characterization is shown in Figure 3.7. The pulse widths range from 3\(\text{ns}\) to 10\(\mu\text{s}\), and the voltage depends on the operable range per pulse width but is typically from 0V to 1.5V.

Therefore the pulsed reset flow becomes as follows:

1. Choose the longest pulse width (time)
2. Choose the smallest pulse height (voltage)
3. Perform 2 ‘set’/‘reset’ sweeps to check for filament stability
4. Apply pulse for reset
5. Perform sweep ‘set’ to check the resulting HRS and reset the filament to LRS
6. Do the pulse ‘reset’ + sweep set (steps 4 and 5) for 10 time to get statistics
7. Advance to next pulse height (step 6)
8. Once done with all voltages advance to next width and repeat from step 1
The complete pulsed reset characterization is illustrated in Figure 3.8. For each pulse width a sweep of pulse voltages can be seen. The approximate initial LRS level and 10X window level for HRS are also indicated. A wealth of information is hidden in this reset characterization.
Some of the major points worth noting are:

1. The maximum pulse heights reached at all these relatively short pulse lengths are higher than achievable by using the standard sweep voltage reset with millisecond time scale.

2. Shorter pulses require higher voltage to achieve same reset.

3. The effect of longer pulses for better reset at a given voltage saturates after a certain pulse width.

4. Short 3ns pulse widths can produce a much better reset than seen with other pulses, albeit with higher pulse voltage.
3.3 Voltage vs. Time Trade-off and Deep Reset Phenomenon

The pulsed reset characterization of the HfO$_2$ RRAM device reported above, quantifies the trade-off between required pulse voltage and pulse length. Figure 3.9 presents the result of vertical slices at three different voltage values from Figure 3.8. Clearly, a linear increase in applied voltage at any given pulse time, produces an exponential increase in HRS resistance.

![Pulsed Reset](image)

Figure 3.9: Impact of pulse time on HRS at fixed pulse voltages. This plot represents vertical slices through the pulsed characterization data presented in Figure 3.8.

We can also see the data from the full pulsed reset characterization as the pulse height and width needed to achieve a similar reset of 10X resistance increase over LRS. Figure 3.10 depicts this data, showing how a shorter pulse duration requires higher pulse voltage to achieve a similar reset operation.
Figure 3.10: Pulse length and voltage needed to achieve a 10X HRS/LRS ratio, depicting the voltage-time trade-off.

The last point to note from the pulsed reset characterization, which is in fact most interesting, is the reset achievable with the ultra-short pulse width of 3ns. As seen from Figure 3.8, the voltages levels needed to achieve a decent reset with the 3ns pulse is much higher than longer width pulses. However, the maximum increase in resistance achievable with the 3ns pulse width at higher voltages is over ten times more than that achieved by the longer pulses. We elaborate on this more, in the following sections.

This observation was further investigated and verified on multiple other devices as well. Four different pulse lengths of 10µs, 100ns, 10ns, and 3ns were used to probe the maximum reset achievable by each pulse length. Using the pulsed reset methodology, each pulse width was swept through the range of voltages from 0V to 5V or until the device no longer was able to switch. The result of this experiment is plotted in Figure 3.11, where each color represents a particular pulse width and the
darker color is HRS while the lighter color represents the LRS values.

As expected, longer pulses produce better reset at lower voltages but there is a maximum limit on the pulse voltage the device can tolerate before an irreversible breakdown occurs. For example, figure 3.11 shows two blue curves representing two similar but separate devices on which 10µs pulses were applied. Each point on the curve represents the HRS result of applying one reset pulse. As the pulse voltage was increased, more resistive HRS is obtained. After each pulse, the ‘set’ sweep initializes the device to the LRS state, which is relatively constant and can be seen in the figure in a lighter blue color. In the case of the 10µs width, a maximum voltage of around 1.8V is seen before the device breaks down and enters a conductive state (even more conductive than the LRS). It can be seen in the figure that currents are as high as 100µA, and they would be higher if it were not for the instrument current compliance limiting the current to this level.

Reducing the pulse width of the applied pulses allows for larger voltages to be applied before the irreversible breakdown is observed. The red curves show that the 100ns pulses allow for voltages up to around 2V, 10ns pulses up to around 3V, and ultra-short 3ns pulses allow up to around 5V pulses to be applied. What is significant is that with this increase in maximum pulse voltage allowed, ‘deeper’ reset is observed. For example, comparing the HRS reached by the 3ns pulses with the HRS reached by 10µs pulses, over 10X improvement in HRS resistance in the ultra short case is observed.

The value of the discovery of this ‘deep reset’ phenomenon with shorter pulses,
in addition to suggesting a way to practically achieve better reset on RRAM devices with shorter pulses, is giving the insight to the current understanding of the reset process and its kinetics. The reset process is understood to increase the resistance of the device in LRS state by rupturing the filament created in the ‘forming’ process. This rupture occurs by oxidation of the tip of the filament thus creating a thin oxide barrier to conduction. During the ‘reset’ process, a competition between reset and breakdown seems to be in effect in which if the applied pulse is long enough with a large applied voltage, the produced thin oxide barrier breaks down and if the pulse is short enough but with larger voltage a better reset is observed.

This barrier can reach a breakdown point depending on the applied voltage and applied pulse length, just as in the TDDB in oxides. ‘Deep Reset’ corroborates that the barrier is an oxidation of the filament, and that this oxidation is occurring at the Bottom Electrode (BE) side since the sudden breakdown can best be understood in this case where no gradual disintegration of the barrier is possible due to the BE TiN functioning as a good barrier to oxygen ion migration.

In the next chapter, we present our model for the switching process in HfO$_2$ based RRAM which is corroborated by the physical and electrical characterization presented in this chapter.
Figure 3.11: 'Deep Reset' Phenomenon: with shorter pulses the device can withstand higher pulse voltages. The data shown in this plot was collected for two devices per pulse width condition using the pulse characterization methodology introduced in this chapter. The darker and lighter shades of each color represent the HRS and LRS, respectively. It can be seen that shorter pulses can achieve deeper resets otherwise not possible with longer pulses.
Chapter 4

Kinetics Model of the Reset Process

Although much progress has been made in understanding the physical mechanism of RRAM device operation, by means of characterization as well as first principle simulations, a model linking the atomic-level material properties to device switching characteristics, capable of guiding optimization of processing and operation conditions is still lacking. Moreover, the size of the devices and the actual operating region of the device, as well as the random nature of atomic ordering demands for a detailed model capable of capturing and simulating the physics based intricacies of RRAM resistance change.

In this chapter, a detailed model for the kinetics of switching in HfO$_2$-based RRAM is presented. Subsequently, based on the filament characteristics as observed by TEM
(see Chapter 3), the kinetics of the reset operation is simulated using the Time Dependent Monte Carlo (TDMC) method incorporating \textit{ab-initio} calculated microscopic characteristics of the oxygen ions in hafnia. Temperature and field driven oxygen diffusion in the oxide surrounding the filament is shown to play an important role in the reset and in providing the needed supply of oxygen to re-oxidize the tip of the Hf-rich metallic filament, thus switching the device to the High Resistance State (HRS). This model points towards the critical reset parameters facilitating the transition to the HRS and ultimately improving device variability.

4.1 Modeling the Forming and Switching Processes

The general device operation of a typical HfO$_2$-based RRAM cell starts with the forming step. As we have seen, this forming process involves applying a sufficient bias across the dielectric resulting in the formation of a conductive filament which exhibits Ohmic-like conduction as observed from linear I-V relation as well as increasing resistance with temperature. This indicates that the conducting filament is consists of a Hf-rich and oxygen deficient region in the dielectric. Thus this forming step consists of oxygen expulsion from the dielectric region to put the device in the LRS.

4.1.1 The Role of Grain Boundaries

A reported study on morphological properties of the HfO$_2$ film assisting the forming process indicates that higher conduction in the oxide region is associated with
grain boundary regions in the oxide. In this report, conductive atomic force microscopy (C-AFM) was employed in which a nanometer resolved characterization of the electrical and topographical features allowed to distinguish individual nanocrystals in from the oxide surface. Results indicate that the current through the dielectric prefer to flow along grain boundaries, and moreover, that the breakdown induced by the applied voltage preferentially occurs at grain boundary sites [66–70].

These experimental observations can be explained by *ab initio* calculation of the grain boundaries (GB) in monolithic HfO$_2$ which indicate that the segregation of oxygen vacancies to the GB are in fact thermodynamically favorable [71]. The higher vacancy densities in the GB leads to creation of a conductive sub-band in the GB region as a result of the overlapping of the localized Hf d-states in the HfO$_2$ band gap. This is responsible for creation the higher conduction percolation paths for the current flow in the fresh devices which initiate the forming process. More details on the mechanism of forming itself will be explained in the next sections.

### 4.1.2 Mechanism of Forming Process

As describe above, the grain boundaries in the HfO$_2$ RRAM device function as initial seeds for potential filament growth. GBs drive the leakage current through the conduction sub-band created by the vacancies that have been created near the GB or diffused near the GB and preferred thermodynamically to stay there. The electric current through the GBs can be explained by Trap Assisted Tunneling (TAT) of electrons that are captured and emitted from these vacancy states. The generation of
new defect states governed by the Equation 2.1 can explain the further growth of these initial GBs into full fledged conductive filaments when an increase in the externally applied voltage or temperature increases the generation rate of new defects.

The forming process driven by an applied external voltage lends to a runaway process that needs to be limited to control the produced filament size to functional dimensions. Various factors lead to this rapid decrease in the device resistance termed the 'runaway' process which are describe below:

1. The electron capture and emission leading to the TAT current through the device is associated with the displacement of the surrounding lattice atoms or lattice relaxation. This lattice relaxation leads to energy dissipation in the form of phonons near the trap location leading to a local temperature increase. The increase in local temperature in turn increases the generation rate as described by Equation 2.1 thus leading to further defect generation near the current defect site. Thus allowing for a positive feedback process of self-accelerated bond breakage leading to the runaway breakdown.

2. Any defect generated in the already conducting percolating path in the GB will lead to a more conductive subsection of the GB, creating a redistribution of the voltage drop across the device so that the less conducting section of the GB now effectively sees a larger voltage. This local increase in voltage leads to higher generation rates and thus a speedup in the filament formation process.

3. Finally, and especially once the forming runaway phase has begun, any decrease in resistance of the device due to defect generation leads to an increase in the
current density and thus Joule Heating in the filament region again leading to higher generation rate of defects in the filament region.

Upon the breakage of the Hf-O bond, the oxygen ions either diffuse outward from the high current region, following the temperature and ion density gradient, or recombine with nearby vacancies. Therefore, the creation of the Hf-rich conductive filament (CF) in the forming process consists of a combination of these three processes of generation, recombination and out-diffusion of oxygen ions from the GB region. Experimentally, this whole process occurs in the very short (sub)nanosecond instance of breakdown observed as the sudden current increase (resistance decrease).

The final shape and dimensions of the CF depends on the precise kinetics of oxygen ion release and out-diffusion into the surrounding oxide region. Once the Hf-O bond breaks, the O\(^{2-}\) ion shifts to an interstitial lattice position leaving behind a vacancy V\(^{2+}\). This O\(^{2-}\) and V\(^{2+}\) pair form a Frenkel pair due to their Coulomb coupling which will recombine if no strong electric field separates them or if no electron is immediately captured to neutralize the vacancy. The breakdown process, however, can proceed with less recombinations if the Frenkel pair is generated near an already neutralized vacancy as the electron can transfer to the newly created vacancy and compensate its positive charge \[42\].

When a cluster of overlapping vacancies has been formed, the injected electrons captured by a cluster are localized throughout the entire cluster volume and can compensate for positive charges associated with oxygen vacancies. Any new vacancy created at the boundary of a vacancy cluster is instantaneously neutralized by the
delocalized electrons in the cluster and so the \( \text{O}^{2-} \) ion is free to diffuse. This also reduces the activation energy of defect generation near vacancy clusters compared to the formation of a vacancy in bulk HfO\(_2\) \cite{72}.

The described model predicts that the created filament would be conical in shape with the wider side being away from the cathode of the device (Figure 4.1). One reason for this is that more energy is dissipated during the electron trapping when the trapping site is farthest away from the cathode due to the acceleration of the injected electron by the applied electric field. This higher power dissipation results in higher rates of oxygen ion generation and out-diffusion from the filament region. Additionally, the anode side of the device is fabricated to contain more vacancies to begin with. This has shown to be key in producing devices capable of switching after the forming step \cite{73}. As we have seen, clusters of vacancies lower the activation energy for new Hf-O bond breakage and thus this region of high vacancy concentration is expected to give room for the larger end of the CF. The physical characterization of the RRAM cells presented in Chapter 3 corroborate this prediction of the conical CF with the larger end being towards the anode.

The properties of the initial conductive filament created during the forming process essentially determine the device switching characteristics that follow. The size and shape of the created filament, the oxygen distribution post-forming, the external electrical and thermal stimulants as well as the device structure all have impact on how and whether a 'reset' to the HRS state is possible. In fact, technically, any oxide can be 'formed', but not any 'formed' device can be 'reset' and thus lead to switching. Contrary to the 'forming' step, the reset process is a less understood process with
Figure 4.1: A schematic of the switching kinetics. During the forming process (left) ionized oxygen atoms out-diffuse and eventually get stored in the interstitial positions around the filament region. During the reset process (right), the O-ions diffuse following the electric field and density gradient and may re-oxidize a portion of the filament rupturing it into the high resistance state (HRS).

much controversy and various conflicting models trying to explain this key process in RRAM switching [1, 55–65]. In the following sections, we study the reset process in detail and propose a physics based model based on \textit{ab initio} calculated parameters to simulate the reset process.

### 4.1.3 The Reset Process

The reset process in HfO$_2$-based RRAM occurs by applying a voltage of opposite polarity from the voltage applied during ‘forming’ or ‘set’. The device is initially in the metallic-like conductive LRS state and the applied field is in the opposite direction of that applied during ‘forming’. The Hf-rich CF is surrounded by oxygen ions released during ‘forming’ that now remain in interstitial locations in the HfO$_2$ lattice. The current flowing through the CF is a source of Joule heating in the CF region and results in a temperature increase in the filament area. The applied electric field and
the increased local temperature increase the diffusion rate of the $O^{−2}$ ion surrounding the CF. The applied field is such that the $O^{−2}$ ions are expected to diffuse towards the bottom electrode of the device where the tip of the conical filament is predicted to be narrower. The increase in the oxygen ion density near the CF leads to recombination of the $O^{−2}$ ions with the existing vacancies in the CF and thus re-oxidation of the CF, in other words, the rupture of the filament. Figure 4.1 depicts a schematic of this described reset process.

Various forces act together to enable the reset process. In the following sections we study each of the major forces and culminate the chapter by simulating these forces together to simulate the reset process. The following are the forces considered in this chapter:
4.2 3D Time Dependent Monte Carlo (TDMC) Simulation of Switching

The kinetics of ‘forming’ and subsequent switching with the described breakdown, diffusion, and re-oxidation model cannot be captured using a closed analytic form due to their probabilistic nature and therefore require Monte Carlo simulations using rate equations, as we will describe in detail [55, 74, 75].

4.2.1 The Diffusion Process

Ab initio calculations suggest that released O\(^{2-}\) ions in HfO\(_2\) diffuse with an activation energy of \(E_A=0.7\text{V}\) by substituting lattice oxygen atoms in the oxide. We can model this diffusion using the Arrhenius rate equation with \(E_A\) as the activation energy as follows:

\[
R_{\text{Diffusion}} = \nu \cdot e^{-\frac{(E_A - q\frac{1}{2}E)}{kT}}
\]  

(4.1)
where, $\nu$ is the lattice vibration frequency and is set to $10^{12}\text{s}^{-1}$, $\lambda$ is the jump distance and is equal to $2.5 \times 10^{-10}\text{m}$, $Q$ is the charge of the diffusion specie ($O^-$), and $\vec{E}$ is local electric field. The local $\vec{E}$ is a superposition of the applied electric field at any position which in turn depends on the material and shape of the local region, and the Coulomb forces experiences by the particle of interest. This Coulomb force is a superposition of the forces from all the particles in the 3 dimensional area surrounding the particle which have non-negligible magnitude. The following equation demonstrates the 3D vector form of the Coulomb field calculation:

$$
\vec{E}_{Coulomb} = \frac{1}{4\pi\epsilon_0 \epsilon_r} \sum_{k=1}^{n} \frac{Q_k(\vec{r} - \vec{r}_k)}{(\vec{r} - \vec{r}_k)^3} \quad (4.2)
$$

where $Q_k$ is the charge of particle $k$, and $\vec{r} - \vec{r}_k$ is the vector from the particle of interest to particle $k$. The total local electric field is then:

$$
\vec{E} = \vec{E}_{Applied} + \vec{E}_{Coulomb} \quad (4.3)
$$

As seen from Equation 4.1, the effect of the total local electric field is to change the effective barrier to diffusion. Thus if the field is in the opposite direction of diffusion, for a negatively charged ion, the barrier to diffusion is reduced and diffusion is more likely to occur (as depicted in Figure 4.3).

Oxygen ions released by the forming process reside in interstitial positions in the HfO$_2$ lattice. As Figure 4.4 demonstrates, the diffusion of this interstitial ion is most
efficiently done by substituting a oxygen atom in the lattice and forcing the lattice atom into a new interstitial position. The electric field can help reduce this barrier to diffusion. The Figure also shows that if the oxygen ion is near a CF, then occupying a CF vacancy site is energetically favorable and has a lower barrier to diffusion.

The diffusion rate given by Equation 4.1 also has a strong temperature dependence which greatly affects the diffusion kinetics of $O^{-2}$ ions according to the local temperature. Capturing the exact kinetics of resistance switching thus requires a precise map of the local temperature updated with changes in conditions, so that accurate diffusion rates can be used and accurate switching times obtained.
4.2.2 Local Temperature and Electrics Field Calculation

The evolution of the shape of the conductive filament (CF), dictates the effective resistance of the memory cell and therefore the current, temperature and electric field locally at different points in the device. The temperature in the CF when the device is in the LRS state and prior to any reset occurring, can be calculated by considering the heating to be purely Joule heating. Determining initial conditions for the CF size and shape as well as the resistivity of the CF Hf-rich metallic material
becomes critical. We can start by assuming that the CF is composed of pure HF metal with resistivity equal to that if bulk Hafnium ($\sim 300 \, n\cdot\Omega\cdot m$). Figure 4.5 depicts the relationship between filament diameter and LRS resistance under these assumptions. Typical LRS values observed range between $1K\Omega$ to $10K\Omega$ corresponding to $5\, \AA$ to $15\, \AA$ diameter size of a CF with cylindrical shape.

![Figure 4.5: Relation of filament diameter to LRS resistance assuming bulk Hf conductivity and cylindrical filament.](image)

Using initial CF size as obtained from our TEM study presented in Chapter 3, as approximately $2.5\, nm$ bottom radius and $5\, nm$ top radius, leads to an LRS CF material resistivity of about $30X$ larger than pure bulk Hf. This filament material resistivity is corroborated by a study in which multiple devices at various different compliance current limits are ‘formed’ leading to different expected initial CF sizes [16]. If we assume the reset is caused by an approximately $1\, nm$ oxide barrier being created on
one size of the filament, and assuming the HRS current is driven by the trap-assisted-
tunneling through this 1nm thick barrier, the exact size of the conductive filament
leading to such HRS current can be obtained and thus the resistivity of the CF in
the LRS state determined.

Thus we perform our simulations with the initial cylindrical filament size of 2.5nm
bottom radius and 5nm top radius with a CF material resistivity of 10 \( \mu \cdot \Omega \cdot \text{m} \) is used.

Joule heating in the CF is then calculated using the following equations:

\[
pC_P \frac{\partial T}{\partial t} - \nabla (k \nabla T) = Q
\]

\[
Q = J \cdot E
\]

\[
J = \frac{1}{\rho(T)} E
\]

\[
E = -\nabla V
\]

\[
\rho(T) = \rho_0 [1 + \alpha(T - T_0)]
\]  \hspace{1cm} (4.4)

where the first equation is the Heat equation, followed by the equation to calcu-
late resistive heating and the equation for linear dependence of resistivity on local
temperature. In these equations, \( T \) is the local temperature, \( k \) is the thermal conduc-
tivity, \( C_P \) is the specific heat capacity, \( Q \) is the heating term, \( p \) is the density, \( V \) is the
voltage drop and \( \rho \) is the electrical resistivity. These equations are solved numerically
using COMSOL Multi-physics on structures designed to reflect the shape and size of
the CF as described above. As seen in these equations, the material resistivity for
the CF can be approximated to have a linear temperature dependence with constant
\[ \alpha = 0.0037 \] or more accurately follow the relation plotted in Figure 4.6. We use the latter in our simulations.

![Temperature Dependence of Electrical Conductivity in Filament 'Hafnium'](image)

Figure 4.6: Time dependence of electrical conductivity of filament 'hafnium-like' material used for simulation.

We simulate the Joule heating with the given initial conditions and CF shape and size. Other than properties modified for the CF material, bulk material properties are used for all other materials in the device. Figure 4.7 shows the structure of the device simulated, as well as the temperature profile obtained from applying a -1.5V reset voltage on the top electrode (TE). The simulations are done in an axisymmetric method in which the resulting profile is assumed to be equivalent when rotated around the y-axis. It is observed that temperatures as high as 1000\(^\circ\)K is reached with this applied voltage, and that this maximum temperature is reached away from the edges of the filament but closer to the bottom Electrode (BE).
Figure 4.7: Temperature profile of device in LRS state right before reset starts.

In the following sections of this chapter we will see how $O^{2-}$ will diffuse into CF, oxidizing the Hf-rich CF, and thus leading to a rupture of the CF. Before any rupture of the CF, and while an Ohmic connection from TE to BE still exists, this Joule heating simulation can be performed. However, once the CF is ruptured, the simulated current flow is disrupted and temperatures are dropped to the ambient temperature, effectively terminating further progress in the simulation. This is while in reality tunneling through the thin barriers formed can still be significant and lead to further filament re-oxidation. The tunneling mechanism for this thin barrier formed near the BE in the CF is mainly assumed to be Direct Tunneling (DT) since the oxide is very thin and Trap-Assisted-Tunneling (TAT) is relatively negligible.
In our structure in COMSOL, we implement this thin barriers as lattice wide rectangles whose height is the tunneling thickness. Figure 4.8 shows a sketch of this tunneling barrier implementation in COMSOL. Each tunneling barrier is allowed to have it’s own independent thickness depending on how the oxygen ions diffuse and re-oxidize the CF. The DT current for each barrier is calculated individually with the heating generated at the interface between the tunneling barrier and the CF where the electrons thermalize after tunneling through the thin oxide. This is modeled by inserting a tunneling resistance where the voltage drops across at the CF and barrier interface. This tunneling resistance generates a tunneling current as calculated from the thickness of the oxide barrier and after which the Heat Equation (Eq. 4.4) is solved to find the temperature distribution.

Figure 4.8: Sketch of tunneling barrier implementation in COMSOL Multi-physics simulator.

The local applied electric field throughout the device is similarly calculated taking
into account the current shape of the CF. Poisson’s equation is numerically solved using COMSOL with material properties as mentioned above. Figure 4.9 shows the electric field distribution for the case of a ruptured CF with a thin oxide barrier formed near the BE side of the filament with a structure as shown in Figure 4.8.

Figure 4.9: Example of the electric field simulations for a filament of 5nm top and 2.5nm bottom diameter with a reset voltage of -1V applied. Poisson’s equation is solved to obtain the electric field shown.
4.2.3 Barrier Creation and Tunneling Mechanism

Electron Tunneling Through a Nanoscale Potential Barrier Between Two Metal Electrodes

As we have seen in previous sections, the diffusion of interstitially stored oxygen ions into the CF and reoxidation can cause a rupture of the filament at which point current tunneling through the created barrier becomes significant. We briefly discuss the direct tunneling calculations used in this work.

The phenomenon of electron tunneling describes the penetration of an electron through a potential barrier higher than the kinetic energy of that electron. There is a long history of work and numerous efforts on describing and calculating the electron tunneling phenomenon.

Esaki’s [76] tunnel diode and his discovery in 1957 was a breakthrough in electron tunneling in solids. Josephson improved significantly the theoretical work in connection with tunneling of electrons between two semiconductors separated by a thin layer of insulating oxide. The well known theory of electron tunneling was developed by Fowler and Nordheim [77] after Richardson’s [78] work on tunneling in which he developed the governing equations from experimental results. Later in 1963 Simmons [79] suggested a generalized formula for the electric tunnel effect covering high and low voltage ranges. Simmons described a method for the application of the theory to a rectangular barrier with the image potential included.
The theory states that if two metallic electrodes are separated by a thin film insulator that acts as a potential barrier, the electronic current can flow from one electrode to the other through the potential barrier. This happens if the electrons of the electrodes have high enough energy to overcome the potential barrier and flow in the conduction band or the gap which separated the two electrodes is thin enough too allow the penetration of electrons.

If \( \phi(x) \) is used to show the potential barrier measured from the Fermi level, the probability \( D(x) \) of penetration of electrons through the potential barrier \( W \) is given by the WBK approximation (refer to Figure 4.10):

\[
D(x) \simeq \exp(-A\sqrt{\phi - W})
\]

\[
A = \left( \frac{4\pi t}{\hbar} \right) \sqrt{2m}
\]

\[
W = E_x - \mu
\]

(4.5)

Where \( E_x = \frac{mv_x^2}{2} \) is the kinetic energy of the electrons and \( t = s_2 - s_1 \). Simmons shows that the number of electrons traveling from Electrode 1 to Electrode 2 is given by:

\[
N_1 = \frac{4\pi m^2}{\hbar^3} \int_0^{E_m} D(E_x) dE_x \int_0^\infty f(E) dE_r
\]

(4.6)

where the maximum energy of the electrons is \( E_m \), the energy in polar coordinates.
is \( E_r \), and \( f(E) \) is Fermi-Dirac distribution function. In a similar fashion, the number of electrons traveling from Electrode 2, which is at positive potential \( V \), to Electrode 1 is given by:

\[
N_2 = \frac{4\pi m^2}{h^3} \int_0^{E_m} D(E_x) dE_x \int_0^\infty f(E + eV) dE_r \tag{4.7}
\]

Subtracting (4.7) from (4.6) and integrating leads to the final current density formula suggested by Simmons:
\[ J = J_0 \left\{ \bar{\phi} e^{-A \bar{\phi}^2} - (\bar{\phi} + eV)e^{-(A(\bar{\phi} + eV)^{1/2})} \right\} \]

\[ J_0 = \frac{e}{2\pi \hbar t^2} \]

\[ A = \left( \frac{4\pi t}{\hbar} \right)(2m)^{1/2} \]

We use Equation 4.8 for calculating the direct tunneling current through the barrier created after the rupture of the CF in the ‘reset’ process. The tunneling current us used to calculate the local temperature and electric field map in the simulated device step by step as the CF changes shapes due to oxygen ion diffusion as simulated in our Kinetic Monte Carlo simulations discussed in the next section.
4.2.4 TDMC Simulation Flow

After discussing the building blocks of the simulation, we now discuss the details of the Time Dependent Monte Carlo simulation implementation. The simulations start with the construction of a 3D unit memory cell containing the multilayer stack of metal electrodes and dielectric material. In addition, for the ‘reset’ process, we start with the initial conductive filament (CF) produced by the ‘forming’ process, as discussed previously, around an initial grain boundary.

This memory unit cell is discretized into a grid of elementary sub-cells with unit length of 2.5Å equivalent to the lattice size of the HfO$_2$ cells. Although the structure is initially symmetric, and temperature and electric fields can be simulated in 2D and mapped to 3D, however, the oxygen ion density and evolution of the ion distribution through coulomb interaction and probabilistic diffusion can only be captured accurately through 3D simulations.

A user defined post-forming interstitial oxygen distribution (see Figure 4.11) is used as the initial condition for the TDMC simulation of oxygen diffusion during the reset process, with each O-ion having a direction-dependent probability of diffusion as described by the hoping rate equation (Equation 4.1). The Joule heating process is solved numerically using Poisson’s equation coupled with the Fourier heat flow equation using COMSOL [80] to obtain the temperature and electric fields in the memory cell region. As discussed, Coulomb interaction of the O$^{2-}$ ions also impacts the hopping barrier and is calculated to be about 0.23eV for the ions spaced at 2.5Å apart as calculated by Equation 4.2.
A time step is defined on the order of the fastest diffusion rate present in the structure and then the simulation progresses by performing the TDMC diffusion with this small time-step. Every oxygen ion can jump in 6 different direction including the diagonals and thus had 22 degrees of freedom and consequently 22 different rates. Multiplying the diffusion rates for every interstitial oxygen ion by the defined time-step produces the hopping probability for that particular ion in that particular direction for the given time-step. If the time-step is too long, then the probability of diffusion increases and can become even a definite jump. Since the ion diffusion is a random process and no ion takes priority over the other, we select the ions in a random fashion and based on the hopping probability decide whether to advance the ion to a new location (assuming the new location is not occupied), or not.

If an ion jumps into a CF region, it is immediately oxidized and the CF becomes oxide in which other ions can interstitially be stored in. After all the oxygen ions
have a chance to advance, the filament shape is updated and a new temperature and electric field is calculated iterating this process during the total applied pulse time. Figure 4.12 shows a flow diagram of this TDMC simulation iteration process described.

Simulation results indicate that the vertical electric field pushes ions towards the bottom electrode, at which point ion diffusion due to density gradient and mutual Coulomb repulsion leads to O-ions moving into and oxidizing the narrow end of the filament. The speed of the process and amount of O-ions available around the filament tip depend strongly on the applied reset voltage. The simulations suggest that higher reset voltage should lead to re-oxidation of a larger section of the filament resulting in a less conductive HRS, as reported experimentally in the literature [15].
CHAPTER 4. KINETICS MODEL OF THE RESET PROCESS

Figure 4.12: Time Dependent Monte Carlo reset simulation flow.
4.3 Simulation of Reset Conditions

The Time Domain Monte Carlo (TDMC) simulations with the given description can be used to characterize the impact of various parameters on the reset process in HfO$_2$ based RRAM. It is important to note that in these simulations we have essentially defined the driving forces and we obtain the results of the simulation of these driving forces. Any parameters used in defining the driving forces are obtained from ab-initio simulations with no fitting parameters. Using this tool we study the effect of pulse length, pulse height, initial oxygen ion concentration, as well as a study of variability in the reset process.

The reset process is essentially a function of the oxygen ion evolution in the RRAM cell. Figure 4.13 shows the progression of the 3D oxygen ion distribution during the reset process. It is observed that, starting from an initially random oxygen ion distribution, the O$^{2-}$ ions tend to disperse due to interaction with neighboring ions and their Coulomb field. Moreover, the ions tend to diffuse toward the BE area due to the applied electric field during reset. The crystalline TiN BE is a good barrier to oxygen and causes the diffused ion to gather at the BE interface. The excess ions at the BE interface forces some ions to push towards the filament and oxidize the tip of the filament thus rupturing it and causing a reset to occur.
Figure 4.13: Two snapshots of the reset simulation progression are displayed. 1) the initial condition for O-ion distribution obtained from a random distribution of O$^{2-}$ ion simulating out-diffusion during forming 2) the final filament rupture moment, also notice the filament narrowing due to progressing of oxidation along its sidewalls.

4.3.1 Pulse Length

We monitor three figures of merit during the simulation, namely the maximum temperature observed at fixed points in time, the maximum electric field, and the current at the applied voltage of 0.1V which is reproducing a read conditions in the memory device (read current). Figure 4.14 demonstrates the progression of these figures of merit throughout the reset process. A 2-dimensional snapshot of the filament cross section is also available showing the filament narrowing and the barrier creation process. It is observed that the reset process can be divided into four distinct stages. The first stage is the LRS state before any reset occurs as can be seen in Figure 4.14 as indicated by the number 1, the second stage corresponds to the filament narrowing (indicated by the number 2), the third stage corresponds to the onset of the filament rupture (indicated by the number 3), and the last stage is the reset occurring
post filament rupture driven by the tunneling current through the thin oxide barrier (indicated by the number 4).

Figure 4.14: Simulated current, temperature, and electric field simulated as a function of pulse length.

As can be seen from Figure 4.14, the length of the pulse at any fixed pulse voltage is critical in ensuring a proper reset occurs. Assuming a large enough pulse voltage is applied, the pulse length has to be sufficient to allow for oxygen ion diffusion and subsequently rupture of the filament.

4.3.2 Pulse Amplitude

We study the effect of varying pulse voltage on the kinetics of the reset process. Figure 4.15 plots the maximum observed electric field as a function of the simulation time. It is observed that the initial e-field is higher in the case with a larger applied
voltage. At the onset of reset the maximum electric field is on the same order of magnitude however no correlation is seen between the applied pulse voltage and this maximum e-field. Finally, The maximum e-field post filament rupture can be seen to be highest with the higher applied pulse voltage, as expected.

![Reset Simulation at Constant Applied Voltage](image)

**Figure 4.15:** Simulated reset maximum e-field in filament area at various constant applied reset voltages plotted as a function of time.

The temperature evolution is also instructive as depicted in Figure 4.16. In all cases, this peak temperature initially decreases as the filament is in the narrowing stage (stage 2) and starts to increase again right before the onset of the filament rupture, after which the temperature drops drastically corresponding to the current decrease after the filament has ruptured. The final temperature corresponds to the
CHAPTER 4. KINETICS MODEL OF THE RESET PROCESS

93
temperature produced by tunneling current through the thin filament barrier. A
strong correlation between maximum temperature and applied pulse voltage is seen
and this temperature difference is the key factor determining the speed of the reset
process.

![Reset Simulation at Constant Applied Voltage](image)

**Figure 4.16:** Simulated reset maximum temperature in filament area at various con-
stant applied reset voltages plotted as a function of time.

Finally, Figure 4.17 shows the read current at a 0.1V applied voltage. This 0.1V
applied voltage is chosen in order to ‘read’ the state of the memory throughout the
reset process, in a fashion that is comparable across different applied reset conditions
(similar to the pulsed reset characterization method introduced in Chapter 3). It is
observed that the reset starts at an earlier time in the case of higher applied pulse
voltage and with a more abrupt filament rupture when compared to the lower applied voltages. The 0.6V case, for example, shows a filament rupture occurring in stages. This is similar to what we see in experimental characterization of the RRAM memory cell as shown in Chapter 3.

Figure 4.17: Simulated reset current at various constant applied reset voltages plotted as a function of time.

4.3.3 Initial Ion Distribution

Modeling the reset process requires knowing the distribution of the oxygen ions in the dielectric resulting from their out-diffusion during the forming process. This initial oxygen ion concentration turns out to be extremely important in determining
the reset characteristics of the cell. We study this by simulating the reset process for various initial interstitial oxygen concentration levels (Figure 4.18).

Figure 4.18: Three different interstitial oxygen concentration levels prior to reset. Blue region is the filament and the red dots are interstitial oxygen.
The reset simulation results for the three oxygen concentrations are significantly different. The simulations were all performed under the constant applied pulse voltage of 1V. Figure 4.19 shows that the reset occurring in the case of ‘100%’ interstitial oxygen concentration as opposed to ‘25%’ concentration can produce a reset faster by about 4 orders of magnitude. Moreover, the reset is stronger and with a thicker oxide barrier when a larger oxygen concentration is present. This shows that being able to control the initial oxygen concentration through process engineering and the forming process are key in producing efficient resets in the HfO$_2$-based RRAM devices.

Figure 4.19: Simulated read current in filament area at various initial oxygen ion densities plotted as a function of time.
4.3.4 Variability

Finally, we study the impact of the reset process on variability in the high resistance state (HRS). One of the key obstacles remaining for RRAM is the large variability in the HRS and LRS levels, as we have seen in Chapter 2. We simulate this variability in the reset case. Figure 4.20 shows the result of this simulation. As can be seen from Figure 4.20, similar to the treatment before, three distinct regions can be defined.

![Reset Simulation at Constant Applied Voltage](image)

Figure 4.20: Variability simulation of reset process for 7 different cases with identical initial conditions.

The first region is the region of filament narrowing, in which the reset is not significant but variability is also small. The second region is the region immediately
prior to the filament rupture, in which the filament has gotten significantly narrow and to the point of rupture. This region shows the greatest variability from one reset process to the next. It is a result of random diffusion of oxygen ions, and the stage in which most of the resistance change is occurring and thus large variation is observed. The final region is that of the tunneling region in which the reset has mainly completed and any remaining reset is due to tunneling occurring in the thin oxide barrier created in the filament.

A similar treatment is performed by Zhao et al. [2] with experimental data in which they apply a reset pulse and measure the resistance vs pulse number. Figure 4.21 shows a plot of such reset results with a sketch of the filament state as the reset is progressing. As can be seen, region two or ‘linear region’ is a region with most variability and one that should try to be avoided through optimization of reset conditions. If the reset pulse is optimized to put the device into region 3 (or ‘saturation region’) more control on variability can be obtained.
Figure 4.21: Experimental data of reset resistance values as a function of applied reset pulse number. The operating range can be divided into three distinct regions as shown [2].
Chapter 5

Conclusions

5.1 Summary of Contributions

In this thesis various key challenges in HfO$_2$ RRAM devices have been addressed. We started from the forming process and addressed variability caused by this initial key step in the device initialization. We showed how controlling the forming process can lead to better uniformity in switching characteristics and also lower power operation. The devices were then physically and electrically characterized, showing evidences of a filament in the oxide region post-forming using scanning TEM. A pulsed characterization methodology was also introduced and used to characterize the RRAM cell using short voltage pulses leading to the observation of the ‘deep reset’ phenomenon. Finally, a physics based switching model is proposed focusing mainly on the controversial reset process and a Kinetic Monte Carlo simulator is developed.
based on this model to simulate the reset process.

The key contribution in this work can be summarized as follows:

1. Low voltage CVF results in:
   
   - Reduction of both “intrinsic” and “extrinsic” variation of $I_{Max}$
   - Lower $I_{Max}$ that increased LRS and HRS lower power operation
   - Improved HRS uniformity (device-to-device and cycle-to-cycle)

2. Physical characterization evidence namely STEM dark image and EELS spectra for a filament in HfO$_2$ RRAM was obtained.

3. More resistive HRS state (‘Deep reset’) is achieved when ultra-short pulses at larger voltages were used. This was otherwise not possible using longer pulse times.

4. A physics-based microscopic kinetics model was developed for the reset process demonstrating:
   
   - Initial O-ion concentration has critical role on the speed of the reset process. Higher initial concentrations of O-ions result is faster reset operation, as simulated.
   - The reset process voltage and time dependence as observed experimentally by pulse measurements is simulated. It i found that a linear increase in applied voltage at any given pulse time, produces an exponential increase
in HRS resistance. However, a similar change in HRS with fixed voltage can only be achieved with exponential change in pulse duration.

- The reset process can be divided into three distinct regimes based on the duration of the reset pulse applied. The first sub-threshold region is caused by shorter pulses in which filament narrowing occurs but not much is reset and thus variability is not significant. The final region is when the filament has been ruptured which is caused by the longest pulses and current is mainly due to tunneling. The Variability in the reset can be attributed to the linear or saturation reset regime. The linear regime is the regime between sub-threshold and saturation. The reset process is simulated using various pulse durations and it is observed that the linear regime is the most critical in producing HRS variability.

5.2 Outlook and Future Work

The work in this thesis provides a holistic analysis on key challenges for HfO$_2$ RRAM devices. It also establishes a framework for physics based Kinetic Monte Carlo simulation of switching operation with defect generation and oxygen diffusion energies obtained directly from ab-initio simulations of molecular structures. This framework is an powerful tool for studying the impact of various operating conditions and device design optimization on the reset process.

The tunneling description for the thin oxide barrier created in the filament by the reset process requires consideration of trap assisted tunneling for a more accurate
analysis. This becomes especially important for thicker than 1nm barriers as direct tunneling becomes negligible. Using the more realistic multi-phonon TAT current simulation for tunneling through the reset barrier and couple with the TDMC reset simulation would improve the accuracy of the simulation.

Finally, this model has been developed for HfO$_2$ RRAM devices. A similar approach can be used to develop models for various material systems and using the KMC simulation physics based simulations can be used to predict switching mechanisms in other material systems.
Bibliography


[46] Yu-Sheng Chen, Wen-Hsing Liu, Heng-Yuan Lee, Pang-Shiu Chen, Sum-Min Wang, Chen-Han Tsai, Yen-Ya Hsu, Pei-Yi Gu, Wei-Su Chen, Frederick Chen, and others. Impact of compliance current overshoot on high resistance state, memory performance, and device yield of HfOx based resistive memory and


[63] D. Ielmini. Modeling the Universal Set/Reset Characteristics of Bipolar RRAM


