CHARACTERISTICS AND APPLICATIONS OF NON-VOLATILE RESISTIVE SWITCHING (MEMRISTOR) DEVICE

by

Shinhyun Choi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering)
in The University of Michigan
2015

Doctoral Committee:

Associate Professor Wei D. Lu, Chair
Professor L. Jay Guo
Professor Cagliyan Kurdak
Associate Professor Zhaohui Zhong

DEDICATION

To my wife Hae-Ryung, my family and friends

ACKNOWLEDGEMENTS

I gratefully acknowledge all the support I've been given during my time at the University of Michigan. I would like to extend special thanks to my mentor Dr. Wei Lu. He has guided me through this process with kindness, patience, and understanding. This dissertation would not have been possible without the many thoughtful conversations we have had over the years. I would also like to thank my committee members, Dr. Cagliyan Kurdak, Dr. L. Gay Guo, and Dr. Zhaohui Zhong for their mentorship and critical assessment of my work, and for generously sharing their knowledge and time.

I would like to thank my parents, family, and friends for supporting me through this challenging quest. A special thank you goes to my wife Hae-Ryung Park for providing me with unwavering love, support, and encouragement. This work would not have been possible without you.

I would like to express my gratitude to past and present members of the laboratory. Thank you to Dr. Siddharth Gaba , Dr. Taeho Moon, Dr. Sungho Kim, Dr. Yuchao Yang, Lin Chen, Patrick Sheridan, Chao Du, Jiantao Zhou, Ugo Otuonye, Jihang Lee, Wen Ma, Fuxi Cai, Jonghoon Shin, and Yeonjoo Jeong for friendship, support, assistance and encouragement.

In addition, I would like to extend my thanks to friends and colleagues in the Electrical Engineering and Computer Science who have provided me with assistance, support, and advice through my graduate career.

I am very thankful and fortunate to have received funding to support my graduate work through a variety of sources. These sources include JeongSong Cultural Foundation (2009 Sep. ~ 2011 Apr.) and SamSung Scholarship (2011 Sep. ~ 2015 Apr.) for financial support and advice.

Table of Contents

DEDICATION .... **ii**

ACKNOWLEDGEMENTS .... **iii**

List of Figures .... **viii**

ABSTRACT .... **xv**

Chapter 1.Introduction **1**

1.1 Non-Volatile Memory **1**

1.2 Flash Memory Market **2**

1.3 Flash Memory and its scaling issue **3**

1.4 Emerging Non-Volatile memories **4**

1.4.1 Phase-change RAM 5

1.4.2 Ferroelectric RAM 7

1.4.3 Magnetic RAM 8

1.5 RRAM **10**

1.5.1 Bipolar Switching vs. Unipolar Switching 12

1.5.2 Cation Migration vs. Anion Migration 13

1.5.3 Memristor 14

1.5.4 Research on RRAM 15

1.6 Organization of the Thesis **16**

1.7 References **18**

Chapter 2.Comprehensive Physical Model of Dynamic Resistive Switching in an Oxide Based RRAM **22**

2.1 Introduction **22**

2.2 Resistive switching behavior of TaO_x/Ta_2O_5 bilayer **23**

2.3 Details of the model **25**

2.4 Conclusion **30**

2.5 References **32**

Chapter 3.Random Telegraph Noise and Resistance Switching Analysis of Oxide Based RRAM **35**

3.1 Introduction **35**

3.2 Device Fabrication and Measurement Setup **36**

3.3 Statistical and Temperature Dependent Studies of RTN **37**

3.4 Electron Transport Experiment in LRS and HRS. **41**

3.5 Analysis based on noise and transport data **45**

3.6 Conclusion **48**

3.7 References **49**

Chapter 4.Retention Failure Experiments and Modeling of Metal-Oxide Based RRAM ... 51
4.1 Introduction ... 51
4.2 Device Fabrication and Measurement Setup ... 51
4.3 Experimental Data ... 53
4.4 Multiphysics Simulation ... 57
4.5 Monotonic Current Increase Behavior ... 59
4.6 Multi-filament Effects ... 60
4.7 Conclusion ... 61
4.8 References ... 62

Chapter 5.Tuning Resistive Switching Characteristics of Tantalum-Oxide RRAM Devices through Si Doping ... 63
5.1 Introduction ... 63
5.2 Device Fabrication ... 64
5.3 Resistive Switching Behavior ... 66
5.4 Switching Dynamics Analysis ... 68
5.5 Evaluation of the Effective Hopping Distance ... 70
5.6 Details of the model ... 72
5.7 *Ab Initio* Study ... 73
5.8 Analog Switching Behavior ... 76
5.9 Conclusion ... 80
5.10 References ... 80

Chapter 6.Data Clustering using RRAM Network ... 84
6.1 Introduction ... 84
6.2 Device Fabrication and Measurement Setup ... 85
6.3 Analog RRAM Behavior ... 86
6.4 Learning in Crossbar Arrays ... 89
6.5 Details of the Training Process ... 93
6.6 Effect of Applied Voltage Amplitude and the Learning Rate ... 96
6.7 The Effect of Device Non-Uniformity ... 97
6.8 Analysis of Performance of the RRAM Network ... 99
6.9 Conclusion ... 100
6.10 Appendix ... 101
6.10.1 Appendix A- Device Modeling ... 101
6.10.2 Appendix B- Normalization of the weights ... 104
6.10.3 Appendix C- Details of the measured data and modelling ... 105
6.11 References ... 107

Chapter 7.Experimental Demonstration of Unsupervised Learning Using RRAM Networks ... 110
7.1 Introduction ... 110
7.2 Device Fabrication and Analog switching behavior ... 110
7.3 Peripheral Circuitry ... 111
7.4 Learning in the RRAM array ... 113

7.5 Data Clustering before and after Applying Unsupervised Learning Rule **116**
7.6 Conclusion **120**
7.7 References **121**

Chapter 8.Summary and Future Work **122**
8.1 Discussion **122**
8.2 Future work **124**
8.2.1 RRAM Crossbar Array for Preprocessing of Neural Signal 124
8.2.2 Device Optimization – Analog Switching 125
8.3 References **126**

List of Figures

Figure 1.1. Applications of non-volatile memories.. 1

Figure 1.2. The demand of non-volatile memories from 2012 to 2017. 2

Figure 1.3. Schematic of floating gate flash memory. R 3

Figure 1.4. (a) Cross-section schematic of the conventional phase-change memory cell. (b) Temperature- applied electrical pulses widths for SET, RESET and Read pulses. 5

Figure 1.5. (a) A schematic circuit diagram for a typical 1T1C FeRAM cell. (b) Polarization-voltage hysteresis of a MFM capacitor. (c) A schematic circuit diagram for a 1T FeFET device. (d) Source-drain current vs. gate voltage hysteresis of a FEFET device. 7

Figure 1.6. a two-terminal switch can be formed with a switching medium sandwiched between a pair of electodes... 10

Figure 1.7. DC I-V characteristics. (a) Digital-like type device. Obtained from stack of Pd/ TaO_x /Ta_2O_5/Pd. (b) Analog-like type device. 11

Figure 1.8. Classification of the switching characteristics in a voltage sweeping experiment. (a) Unipolar switching. (b) Bipolar switching. 12

Figure 2.1. Modeling a tantalum oxide RRAM during set/reset. (a) Schematic and cross-sectional TEM images of the Pd/Ta_2O_5/TaO_x/Pd bilayer RRAM device. (b) Measured and calculated DC *I-V* characteristics of the Pd/Ta_2O_5/TaO_x/Pd device. The measured device size is 50 nm × 50 nm, and the voltage sweep speed is 2 V/s. (c) Calculated 2-D maps of n_D as well as (d) 1-D profiles of n_D along the center of the CF in the initial state, after reset, and after the set process. The depleted gap is determined as the position where $n_D = 5 \times 10^{20}$ cm^{-3}. The $z = 0$ position is the Ta_2O_5/TaO_x interface. 24

Figure 2.2. Equations and parameters in the proposed model. Three PDEs are self-consistently solved with a numerical solver. 26

Figure 2.3. Parameters from measurements and assumptions. (a) Electrical conductivity pre-exponential factor σ_0, (b) assumed activation energy for conduction E_{AC}, and (c) assumed thermal conductivity k_{th} as a function of local V_O density n_D. 28

Figure 2.4. Simulated geometry used in the calculation. The axisymmetric geometry reduces the problem from 3-D to 2-D. A uniform doping concentration of $n_D = 1 \times 10^{21}$ cm^{-3} was assumed within the CF and the TaO_x layer as the initial state right after electroforming. 29

Figure 3.1. (a) I-V characteristics of a typical device showing the bipolar switching effects. Inset: SEM image of the device. Scale bar is 5µm. (b) Current-time plots measured at 0.1 V for LRS and HRS, respectively. Insets: zoomed-in plots of the circled areas for LRS (left) and HRS (right), showing pronounced RTN in HRS. 37

Figure 3.2. (a) Time-domain analysis of the RTN behavior showing raw data (red) and reproduced data (grey) based on the capture program. (b) Histograms of current vs. occurrence showing a bimodal distribution corresponding to the two current levels causing RTN. (c), (d) Histograms of the dwell times in the upper (c) and lower (d) current levels. The red lines are Poisson fits using as the only fitting parameter. 38

Figure 3.3. (a) Temperature dependence of the characteristic dwell times in the upper and lower current levels. The lines are fits following the Arrhenius equation. (b) Schematic of the cause for RTN. The trapping and detrapping of a trap site near the channel leads to jumps in discrete current levels. The dashed circle represents the area that may be electrostatically depleted by the trapped electron. (c),(d) Histograms and fits of the dwell times at the upper current levels at 250 K and 300 K, respectively. 39

Figure 3.4. (a) Temperature dependence of electron transport in LRS. (b) Temperature dependence of electron transport in HRS. Inset: Schematic of the hopping process. (c) Solid line: I-V characteristics without the series resistor, showing switching between HRS and LRS; dashed line: I-V characteristics with a 1 kΩ series resistor. The device is programmed to an intermediate state instead. Inset: The circuit schematic. (d) Temperature dependence of electron transport in the intermediate state. Inset: Hopping with more closely spaced trap sites and lower hopping energy in the intermediate state compared to the HRS. 42

Figure 3.5. (a) Schematics showing the changes in V_O distribution for the HRS, the intermediate state and the LRS in the Ta_2O_5 switching layer, respectively. The dashed lines in (b) and (c) highlight the filament region with higher V_O concentration than the rest of the film. (d-f) Corresponding changes in the overlap of electron wavefunctions lead to different resistance values for the HRS, intermediate state and LRS. dots: localized states, Gray dashed circle: the localization radii. 45

Figure 3.6. Current-time plots measured at 0.1 V on the same device after two different set and reset process. 46

Figure 4.1. (a) Schematic of the $Pd/Ta_2O_5/TaO_x/Pd$ bilayer RRAM device. (b) DC I-V characteristics of the device showing the bipolar switching behavior. Inset: SEM image of the device. Scale bar is 20 μm. (c) A custom-built high temperature measurement setup using a tube furnace. The left part of the tube is connected to a vacuum pump and the right part of the tube is connected to electrical feedthroughs. (d)The wirebonded devices on a chip carrier in the furnace connected to electrical feedthroughs. (e) I-V characteristics of the device in log scale. (f) Retention measurement results at 320 ℃. A read pulse (0.1 V/10 ms) was applied every 6s during the test. 53

Figure 4.2. (a) Temperature dependent retention measurements at 300 ℃, 320 ℃ , 340 ℃ and 360 ℃. (b) Temperature dependence of the characteristic retention failure time (squares) and fitting (line) following the Arrhenius equation. (c) Schematics showing the changes in V_O distribution from the LRS (i), after VO out diffusion (ii), and eventual rupture of the filament (iii). (d) Oxygen vacancy concentration profile predicted from the simple analytical model as a function of time. Dashed red line indicates the critical oxygen vacancies density (defined as $5\times 10^{20}/cm^3$). 55

Figure 4.3. (a) 2-D map of oxygen vacancy concentration obtained through numerical simulations, for in the initial state (LRS). The x=0 position is the center of the conductive filament. (b) Measured and calculated DC I- V characteristics of the device at 320 ℃ showing the model can capture the essential dynamic V_O migration processes. (c) Oxygen vacancy concentration profile calculated from the numerical simulation as a function of time. (d) Measured and simulation results showing the device retention behavior at 320 ℃ inset: Peak V_O concentration at different time instants (A- D). Dashed red line indicates the critical oxygen vacancies density (defined as $5\times 10^{20}/cm^3$). 57

Figure 4.4. (a) 2-D map of oxygen vacancy concentration showing the evolution of the filament at different time scales (corresponding to points A-D in Figure 4b) (b) Measured and calculated device conductance as a function of time at 300 ℃. (c) Effective filament diameter as a function of time. The filament was defined as the region with oxygen vacancy concentration higher than 5×10^{20} cm^{-3}. 59

Figure 4.5. Measured (black line) and calculated (squares) conductance as a function of time at 340 ℃, showing the possible existence of multiple filaments. Evolutions of the two filaments (triangles and circles) were obtained through simulation and the overall conductance (squares) is the sum of the two filaments. 61

Figure 5.1. (a) Conceptual schematic of the oxide RRAM during RS. The agglomerated V_Os enhance the local electrical conductivity and form the CF. (b) Schematic of the potential energy landscape for ion hopping under electric field *E*. (c) Schematic plot of the Pd/Si:Ta_2O_5/TaO_x/Pd bilayer RRAM device. (d) DC *I*–*V* characteristics of undoped Ta_2O_5, $Si_{2.7\%}$:Ta_2O_5, and $Si_{4.2\%}$:Ta_2O_5 devices. The measured device size is 1 μm × 1 μm, and the voltage sweep speed is 1 V/s. 65

Figure 5.2. (a) Switching dynamics characterized by pulse measurements with increasing reset pulse amplitudes. The set-pulse amplitude is fixed at -1.5 V. Before the pulse measurements, the devices are set to LRS with a DC voltage sweep. (b) Schematic of the HRS and LRS for different V_O drift velocities. δ and r increase as v is increased through Si doping. (c) Time-dependent switching transient during the reset and (d) set processes. 68

Figure 5.3. The three-step measurement procedure to evaluate the effective hopping distance. (a) $V_{FORMING}$–d_{oxide} plot to evaluate E_{SET}; the slope represents E_{SET}. (b) V_{SET}–d_{oxide} plot to estimate δ; the y-intercept represents $E_{SET}\cdot\delta$. (c) V_{SET}–ln(t_{SET}) plot to extract a; the slope represents $(2kT/q)\cdot(\delta/a)$. 70

Figure 5.4. Measurement results of (a) $V_{FORMING}$–d_{oxide} to extract E_{SET} and (b) V_{SET}–d_{oxide} to extract δ in the three different samples. 72

Figure 5.5. (a) Snapshots of the amorphous Ta_2O_5 and Si-doped Ta_2O_5 structures obtained in the *ab initio* simulation. The Ta, O, and Si atoms are colored in dark green, red, and blue, respectively. (b) Pair-correlation functions of the amorphous Ta_2O_5 and Si-doped Ta_2O_5 calculated at room temperature. (c) Histograms of the O–O distance from a selected oxygen atom to a neighboring oxygen atom. Three oxygen atoms are selected randomly, as shown in panel a. (d) O and Ta atomic ratio near the selected oxygen atoms. 74

Figure 5.6. (a) Schematic illustration showing a synapse connecting a pair of neurons, where the synaptic functions can be emulated by RRAM devices. (b) Analog switching behavior obtained by pulse trains consisting of 150 reset pulses (1.1 V, 10 μs) followed by 150 set pulses (-0.9 V, 100 μs) with small, nonperturbative read voltage pulses (0.2 V, 1 ms) applied in the intervals. The conductance changes are measured during the read pulse and plotted as a function of applied pulse number. The error bars indicate the standard deviation from the measured data set, which are collected from 50 such test cycles in five different devices in each case. 76

Figure 5.7. Analog switching behaviors obtained by pulse trains in four different cases. 77

Figure 5.8. Implementing four different types of STDP using tantalum oxide RRAM. The pre-spike voltage (V_{Pre}) and post-spike voltage (V_{Post}) are applied to the TE and BE of the RRAM, respectively. The net programming voltage ($V_{Pre} - V_{Post}$) applied across the device depends on the positive or negative moments $t_{Post} - t_{Pre}$. The dots indicate the experimental data, and the lines are guides to the eye. The insets show the (red) pre- and (blue) postsynaptic spike schemes. ... 78

Figure 5.9. Measured cycling endurance performance of analog switching in (a) a Ta_2O_{5-x} RRAM and (b) a $Si_{4.2\%}$:Ta_2O_{5-x} RRAM. Each test cycle consists of a pulse train including 50 reset (1.25 V, 10 µs) pulses followed by 50 set (-1.0 V, 10 µs) pulses. ... 79

Figure 6.1. Modelling the switching performance of a RRAM. (a) DC I-V characteristics of a typical RRAM device showing the bipolar switching. (b) Schematic image of a RRAM device having oxygen vacancy filament. (c) Calculated conductance and internal state variable with 100 pulses of potentiation (-1 V, 10µs) and depression (1.15 V, 10µs), consecutively. (d) The sequences of the applied pulses showing 4 sets of 100 pulses of potentiation and 100 pulses of depression. (e) The measured and calculated conductance changes measured by read (0.2V) pulse with the set and reset processes shown in Fig. 1(d). ... 87

Figure 6.2. The network schematic. The column electrodes represent inputs and the row columns represent outputs. The RRAM devices are located at the intersections where the column electrodes and row electrodes connected. ... 89

Figure 6.3. The result of principal component analysis. (a) The result of read process through RRAM devices showing y_1 at horizontal axis and y_2 at vertical axis before learning process. (b) Principal component analysis using traditional covariance matrix of the data. (c) Principal component analysis using Sanger's rule without the RRAM model. (d) Principal component analysis using Sanger's rule with the RRAM model. ... 93

Figure 6.4. Weights distribution changes for (a) the primary principal component, (b) the second principal component before and after learning process. ... 95

Figure 6.5. Weights changes with individual learning cycles for (a) the primary principal component, (b) the secondary principal component. ... 95

Figure 6.6. The effects of potentiation/depression voltage amplitudes and learning rate changes. (a) The histogram of applied pulse widths as a function of potentiation/depression voltage amplitude. (b) The weight changes as a function of learning rate. ... 97

Figure 6.7. The effect of non-uniformity issue of the devices. (a) The measured data for the analog switching. The blue line and error bar represent the average and standard deviation, respectively. (b) Calculated analog behaviors adding the non-uniformity of the devices. (c) The result of the principal component analysis without device non-uniformity. (d) The result of the principal component analysis with device uniformity. 98

Figure 6.8. Energy barrier of ion hopping process. 101

Figure 6.9. The details of conductance change measured at 0.2 V with 100 pulses of potentiation (-1 V) and 100 pulses of depression (1.15 V), consecutively for (a) measured conductance for 9 RRAM devices of the primary principal component (b) measured conductance for 9 RRAM devices of the second principal component (c) simulated conductance for 9 RRAM devices of the primary principal component (d) calculated conductance for 9 RRAM devices for the secondary principal component. 105

Figure 7.1 Device fabrication and analog switching behavior. (a) SEM images of the fabricated two sets of 16 by 1 RRAM devices. Scale bar: 100 µm. (b) DC I-V characteristics of a typical RRAM device showing the bipolar switching with 100 pulses of potentiation (-1 V, 10µs) and depression (1.15 V, 10µs), consecutively. Inset: schematic image of a RRAM device having oxygen vacancy filament. This is not to scale. 111

Figure 7.2 Peripheral circuitry (a) the photo image of the board with label parts. (b) schematic of the procedure of the board operation...112

Figure 7.3 (a) Experimental measurements collected by the board for 9 RRAM devices in the same column (corresponding to the second principle component), showing the analog conductance change and device-device variations. The conductance was measured with 0.2 V, 1 ms pulses, and the devices were subject to 100 pulses of potentiation (-1 V, 10 µs) and 100 pulses of depression (1.15 V, 10 µs). (b) The solid line and the error bars represent the average and standard deviation. 113

Figure 7.4 Flowchart showing the overall operation procedure. 115

Figure 7.5 Results of principal component analysis. The data are plotted on y_1 and y_2 axis. (a) Initial results of an untrained RRAM network. (b) Results of a partially trained RRAM network. (c) Results of a fully trained RRAM network. 116

Figure 7.6 Weights constituting (a) the primary principal component and (b) the secondary principal component before (upper graph) and after (lower graph) the learning process. 117

Figure 7.7 Evolution of the Euclidean norm of weights during learning. Red dots represents the norm value of the weights for the primary principal component and dark cyan dots shows the norm value of the weights for the secondary principal component. 118

Figure 7.8 Classification based on linear decision boundary (black line) on the clustered data. 119

Figure 8.1. (a) Measured neural signals in motor cortex from a Monkey. Data obtained from Cortical Neural Prosthetics Lab from Biomedical Engineering in University of Michigan. (b) Extracted spikes with constant voltage from Fig. 1(a). 124

ABSTRACT

Non-volatile memory technology scaling has been driven by the ever increasing needs of high-capacity and low-cost data storage. Scaling the conventional floating gate device structure, however, has faced with several technical challenges due to constraints of electrostatics and reliability. Alternative memory approaches based on non-transistor structures has been extensively studied. Among the new approaches, resistive switching devices (RRAM) have attracted tremendous attention due to their high endurance, sub-nanosecond switching, long retention, scalability, low power consumption, high ON/OFF ratio and CMOS compatibility.

In this thesis, we present a systematic study on the fundamental understanding and potential applications of RRAMs. Firstly, we introduce a quantitative and accurate model of the dynamic resistive switching processes, by solving the coupled equations for oxygen vacancy transport, current continuity and Joule heating. Secondly, we show systematic investigations on the resistance switching mechanism through detailed noise and transport analysis, and develop a unified model to explain the conduction path and account for the resistance switching effects. Thirdly, we perform detailed retention studies of oxide-based RRAMs at elevated temperatures and develop an oxygen diffusion reliability model of RRAM devices. The activation energy for oxygen vacancy diffusion is directly calculated from the measurement. Analytical modeling and detailed numerical multi-physics simulation is discussed. Fourthly, we report that doping tantalum oxide based RRAM with silicon atoms leads to larger dynamic ranges with improved accessibility to the intermediate states which is suited for neuromorphic computing applications.

Lastly, we investigate the application of RRAMs in neuromorphic computing by showing data clustering based on unsupervised learning. Through both simulation and experimental studies, we demonstrate that a crossbar array of RRAM devices can perform data clustering through unsupervised learning and enable effective data classification in a real-world problem.

These studies have not only helped the development and optimization of RRAM devices but also highlighted their application potential beyond simple memory. We believe continued development of this emerging device structure may lead to future high-performance and energy efficient memory and logic hardware systems.

Chapter 1.
Introduction

1.1 Non-Volatile Memory

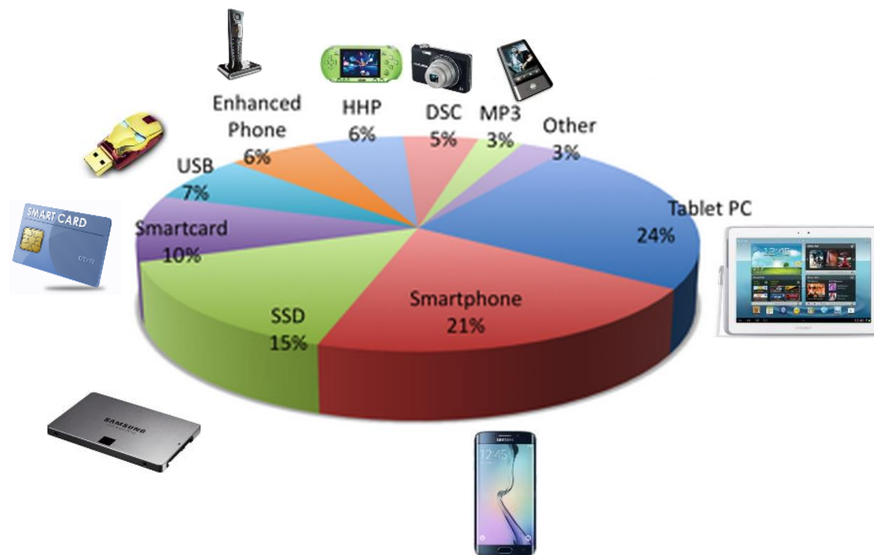


Figure 1.1. Applications of non-volatile memories. Figure adapted and modified from [1].

Non-volatile memories are widely used in many applications such as tablet PCs, smart phones, and solid state drives as shown in figure 1.1. The market of non-volatile memory is rapidly growing and has become a main driver for the semiconductor industry. For example, 38% annual growth is expected from 2013 to 2017, as shown in figure 1.2.

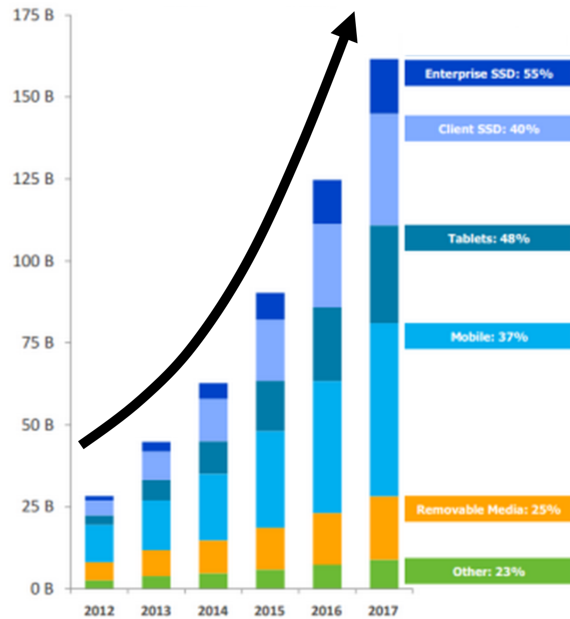


Figure 1.2. The demand of non-volatile memories from 2012 to 2017. Reproduced from [2].

1.2 Flash Memory Market

Flash memory is currently the unquestionable leader of the non-volatile market. The demand of flash memory market has been rapidly increasing because of the advantages such as relatively low cost, small size, and fast speed (compared to hard drives). Flash memory has been a popular choice for the mobile platforms from digital camera to tablets and smart phones. The exponential growth of flash memory is expected to continue due to data explosion from personal communication to large commercial data.

1.3 **Flash Memory and its scaling issue**

A non-volatile flash memory cell based on a floating gate is shown in Fig 1.3. It is based on a normal metal oxide semiconductor (MOS) structure with a floating gate and tunneling oxide. The control gate and the floating gate are made of polysilicon and the tunneling oxide thickness is ~100 Å. The inter poly dielectric layer between the two gates is made of oxide/nitride/oxide (ONO) structure and the thickness is 120 ~140 Å. To program the device, hot electrons in the channel with high kinetic energy can overcome the tunneling oxide/silicon barrier (V_{GS} = ~ 10 V,

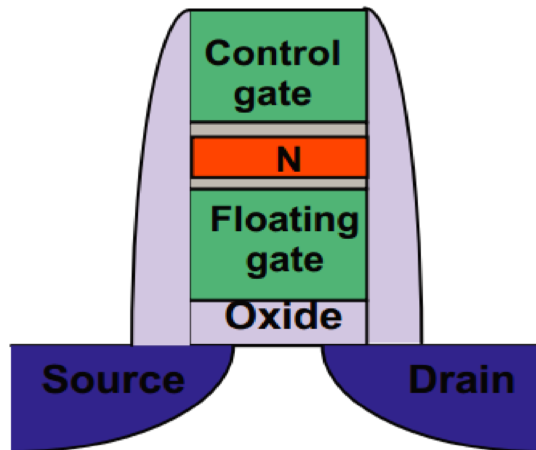


Figure 1.3. Schematic of floating gate flash memory. Reproduced from [3].

V_{DS} = ~ 4 V). The electrons captured in the floating gate modulate the threshold voltage (V_t) of the cell. To erase the device, Fowler-Nordheim (FN) tunneling is utilized. By applying high negative bias (~ -8 V) to the control gate and high positive bias (~ -8V) to the substrate, the electrons are driven back into the silicon layer [4].

Flash memory scaling is driven by the market needs of high-capacity, low-cost data storage capabilities. Scaling the Flash device structure has however encountered several technical challenges [5,6]. First, floating gate interference is the one of the main problems. The space between the floating gate in one cell and the floating gate in an adjacent cell has become small enough that it may cause the charges in one float gate to cause V_t shift in the adjacent cell hence leading to a read error.

Second, tunnel oxide scaling is also a limitation factor. Flash memory utilizes tunneling effect which can cause damage to the tunnel oxide. The operation of flash memory is limited by stress induced leakage current (SILC) related charge transfer problems. Scaling (e.g. reducing the thickness of) the tunnel oxide exacerbates SILC effect.

Third, flash memory scaling increases the impact of single-electron trapping/detrapping in tunnel oxide which causes random telegraph noise (RTN). This results in large V_t instabilities. Additionally, with the scaled flash memory with constant V_t, small number of stored electrons are involved, resulting in large error with the loss of even a few electrons.

Fourth, the reduction of the number of electrons in the floating gate causes few electron phenomena, resulting in stochastic operation in the program, erase, and retention that lead to large variability and unreliable operations.

1.4 Emerging Non-Volatile memories

To address the issues related with Flash memory scaling, alternative memory approaches are being explored. In this section, candidates including Phase-change RAM, Ferroelectric RAM and Magnetoresistive RAM as emerging non-volatile memories are introduced.

1.4.1 Phase-change RAM

Phase-change random access memory (also known as PCRAM, PRAM or Chalcogenide RAM) utilizes the memory effect from the phase change of a material. This material switches back and forth between the amorphous (high resistivity) and crystalline (low resistivity) states. The device structure is shown is Fig.1.4(a). For the phage change material (PCM), a chalcogenide alloy of germanium, antimony and tellurium (GeSbTe or GST) is used [8]. As the device is fabricated, the PCM is usually in the crystalline phase due to the processing temperature. To change the PCM to the amorphous phase, the programming region shown in Fig. 1.4(a) is melted by Joule heating and quenched rapidly to room temperature. As shown in Fig.

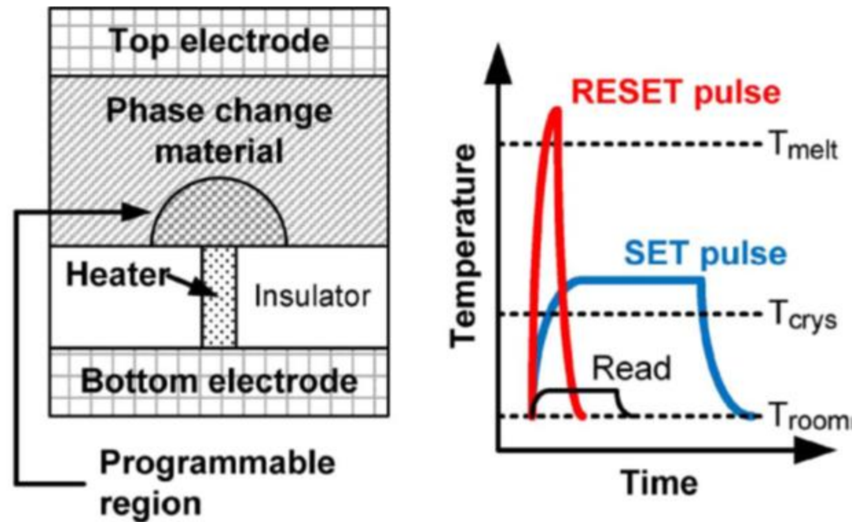


Figure 1.4. (a) Cross-section schematic of the conventional phase-change memory cell. (b) Temperature-applied electrical pulses widths for SET, RESET and Read pulses. Reproduced from [7].

1.4(b), the applied pulse is large to achieve high temperature (600 °C) for a short time. In this process, the PCM loses its crystallinity and stays in an amorphous glass-like state. This amorphous region, along with the remaining crystalline region in the rest of the film in series,

determines the resistance of the cell. On the other hand, to change the PCM back to the crystalline state, a current pulse with medium amplitude is applied to achieve a temperature between the melting point and the crystallization point. Additionally, the pulse is applied for a long enough time (in the order of 100ns) to ensure the film has enough time to fully crystallize. To read the device state (high resistivity or low resistivity), a low read current pulse is applied to keep the device from disturbance as shown in Fig. 1.4(b). However, the continuous heating and quenching process causes void formation (stuck at high resistivity) or elemental segregation. Additionally, thermal coupling between adjacent cells also limits the scaling potential of PCM devices.

1.4.2 Ferroelectric RAM

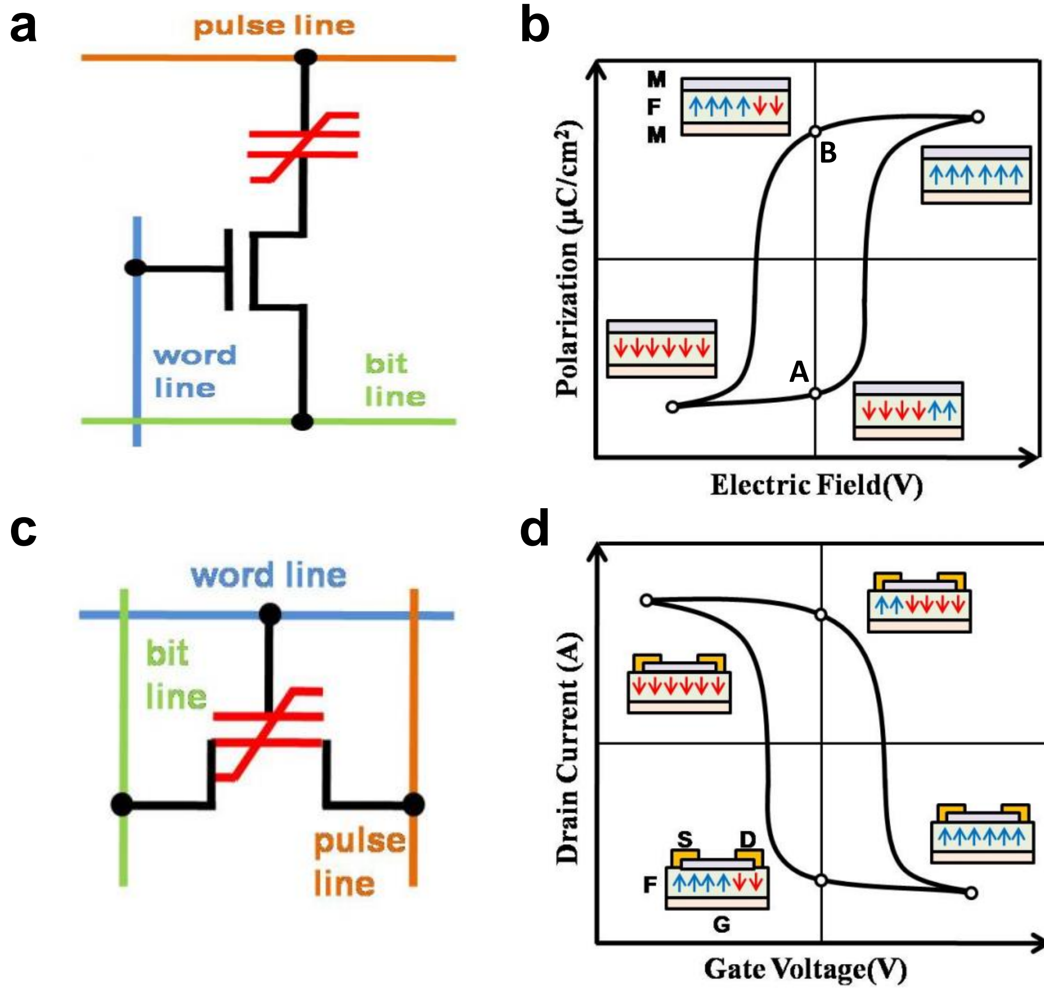


Figure 1.5. (a) A schematic circuit diagram for a typical 1T1C FeRAM cell. (b) Polarization-voltage hysteresis of a MFM capacitor. (c) A schematic circuit diagram for a 1T FeFET device. (d) Source-drain current vs. gate voltage hysteresis of a FEFET device. Reproduced from [9].

Ferroelectric RAM (Ferroelectric random access memory or FeRRAM) is another non-volatile RAM structure. In a conventional 1T1C structure, the device resembles conventional DRAM cells consisting of a transistor structure and a capacitor, whereas the FeRAM utilizes a ferroelectric layer instead of a dielectric layer in the capacitor as shown in Fig. 1.5(a). Typically

lead zirconate titanate (PZT) is used in FeRAM. The polarization properties of a ferroelectric layer are used to achieve non-volatile memory effect. When the external electric field is applied across the ferroelectric layer, the dipole moments inside the ferroelectric layer align to the field direction above the coercive field and leading to a hysteresis effect in the polarization-electric field measuremnets as shown is Fig 1.5(b). After the external electric field is removed, the ferroelectric material retains its polarization as shown in points A and B of Fig. 1.5(b). FeRAM shows low power consumption, fast speed and high endurance. However, the main problem of FeRAM is the destructive read process. The read process utilizes writing process to the cell. If a small (displacement) current pulse is detected, it implies that the device state (polarization) changes and it was OFF state. The low density associated with the capacitor structure is also another bottleneck of a FeRAM.

To overcome the limitation of destructive read process and scaling issues, FeRAM structures without the capacitor (1T structure) have been proposed, as shown in Fig. 1.5(c). This structure utilizes ferroelectric materials as the gate insulator. The polarity of ferroelectric materials modulates the channel conductance of the underlying semiconductor, so the memory effect is obtained as a shift in threshold voltage as shown in Fig. 1.5(d). However, new challenges such as chemical reactions and intermixing between Si and the ferroelectric stack and the short retention time prevent it from challenging the current flash memory market.

1.4.3 Magnetic RAM

Magnetic RAM (MRAM) uses magnetic effects instead of electrical charges currently used in memory technology. A MRAM cell has two ferromagnetic electrodes. While one has a

fixed magnetic polarity, the polarity of the other can be switchable. Between the two layers, there is a thin insulating tunnel barrier. If the polarities of the two electrodes are parallel to each other, the tunnel resistance of the device is low (ON state). On the other hand, if the polarities are anti-parallel to each other, a high tunnel resistance state (OFF state) is obtained [10-12]. To switch the magnetization of the free layer, a large current (on the order of 10mA) is needed to produce the required external magnetic field, which is disadvantage of MRAM. The crosstalk issue at high density due to the spread of the magnetic field into neighboring cells is also a problem, even though MRAM has advantages of high speed and very long endurance.

Alternatively, instead of using an external magnetic field, the magnetization of the free layer can be switched by the spin transfer torque (STT) effect. In this case, a simple two-terminal structure can be used and STT-based memory (STT-MRAM) has attracted significant interest because the switching current in STT-MRAM decreases when the technology scales down. The basic structure consists of one transistor connected in series with the magnetic tunnel junction (MTJ). However, the switching current is still too high for most commercial applications. Additionally, the energy barrier for spontaneous magnetization relaxation is reduced as the magnetic electrode size is scaled, leading to higher error rates.

In the following section, we will discuss resistive memory which has recently emerged as a leading candidate of future non-volatile memory.

1.5 **RRAM**

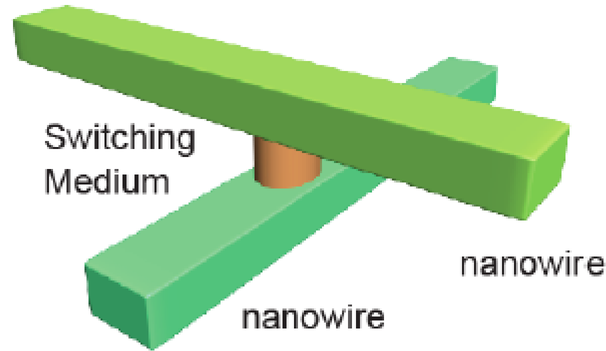


Figure 1.6. a two-terminal switch can be formed with a switching medium sandwiched between a pair of electodes. Reproduced from [13].

RRAM (resistive random access memory) is a two terminal device which consists of a top electrode, a switching medium, and a bottom electrode, as shown in Fig 1.6. The top and bottom electrodes can be metal or compound material with high conductivity. The switching medium sandwiched by the two electrodes is normally an insulating material. This memory cell can be integrated in high-density because a device can be formed at each point in which the top electrode and bottom electrode are crossed to each other. The resistance of the switching medium material can be modulated by applying voltage or current between the electrodes and be reset (high resistivity) and set (low resistivity) repeatedly.

In Fig. 1.7(a), digital-type switching behavior with abrupt resistance changes is shown. This type of device has been utilized for non-volatile memory used for data storage. To date, RRAM has demonstrated endurance up to 10^{12} [14,15], subnanosecond switching [16], device size scaling down to 10nm [17], long retention [18], low energy consumption [19], high ON/OFF ratio [20], 3D structure [21], and CMOS compatibility [22,23].

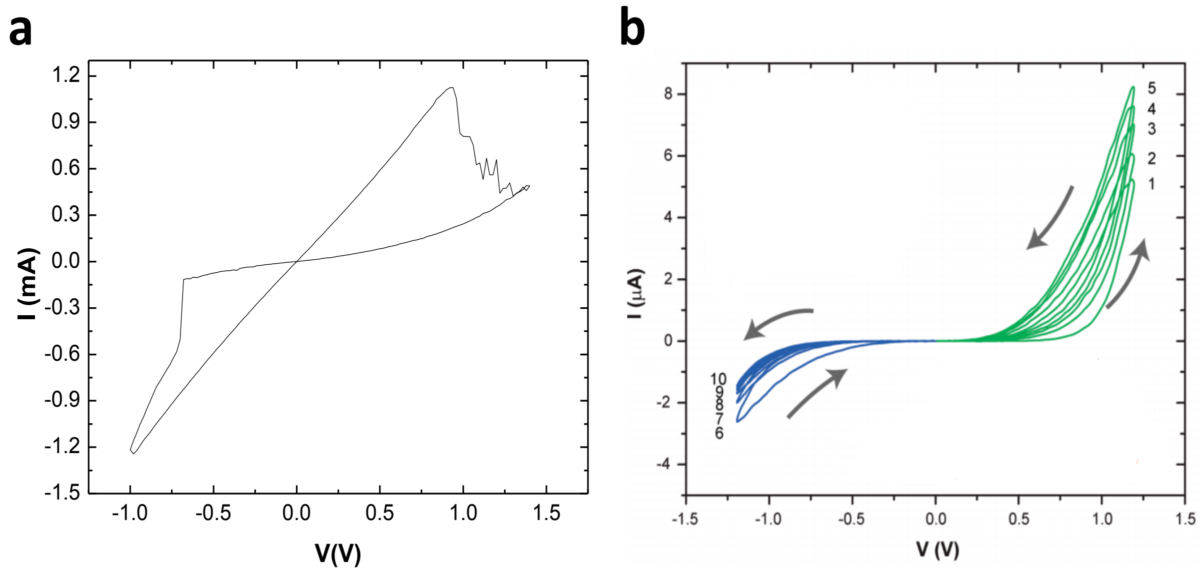


Figure 1.7. DC I-V characteristics. (a) Digital-like type device. Obtained from stack of Pd/ TaO_x /Ta_2O_5/Pd. (b) Analog-like type device. Reproduced from [22].

Figure 1.7(b) illustrates analog-type behavior which has continuous resistance changes. While the digital-type switching has been utilized for non-volatile memory used for data storage, analog-type switching is being investigated for other applications such as neuromorphic computing. The continuous resistance change can mimic the tunable synaptic weight that modulates signals between neurons. This allows synaptic functions to be directly implemented to hardware based neuromorphic systems [25,26]. Several groups including us have demonstrated fundamental synaptic learning functions such as spike-timing-dependent-plasticity (STDP) [27], short-term and long-term plasticity [24,26], and frequency-dependent plasticity [24,28].

There are different ways to classify RRAM devices. The most natural classifications are based on switching characteristics (Unipolar vs. Bipolar) and switching mechanisms (Electrochemical metallization vs. valence change).

1.5.1 Bipolar Switching vs. Unipolar Switching

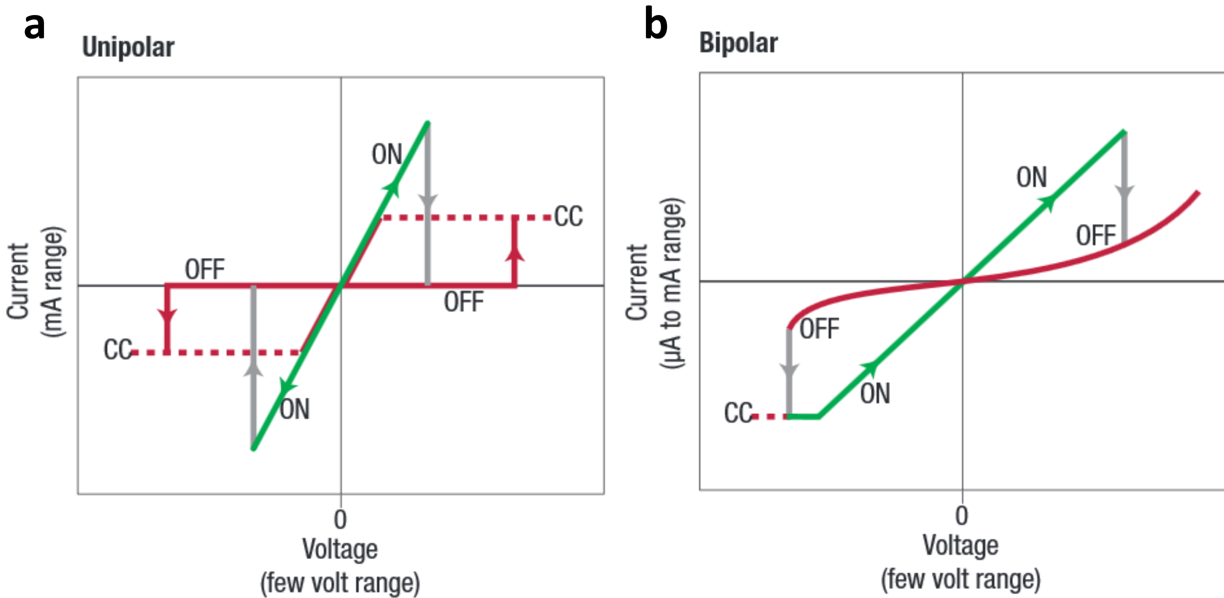


Figure 1.8. Classification of the switching characteristics in a voltage sweeping experiment. (a) Unipolar switching. (b) Bipolar switching. Reproduced from [29].

RRAM devices can be divided into two categories with respect to the electrical polarity required for the switching: bipolar switching device and unipolar switching device as shown in Fig. 1.8. For the unipolar devices, the switching behavior does not rely on the polarity of the applied programming voltage. With a single voltage polarity, the device can switch from ON to OFF or from OFF to ON. As shown in Fig. 1.8(a) the set voltage is always higher than reset voltage. For the unipolar device, Joule heating is believed as the main reason of the resistance switching as conducting filaments can form and rupture inside the switching layer assisted by thermal effects. During the set transition, a partial breakdown occurs in the dielectric layer, followed by conductive filament formation which is modulated by Joule heat, leading to the low-resistance state (LRS). During the reset transition, the filaments are ruptured thermally due to the

high current density as shown in Fig. 1.8(a), leading to the high-resistance state (HRS). For the bipolar devices, the switching behavior depends critically on the polarity of the programming voltage. As shown in Fig. 1.8(b), opposite voltage polarities are required for the reset and the set processes. This implies that the switching mechanism is an electric-field driven process that leads to the formation and rupture of the filaments in the dielectric. More detailed discussions on the switching mechanism will be presented in chapter 2.

1.5.2 Cation Migration vs. Anion Migration

Based on the active species that lead to resistive switching, RRAM devices can be characterized as electrochemical metallization (ECM) or valency change (VCM) devices. Resistive switching devices based on cation migration are called electrochemical metallization (ECM) or conductive bridge random access memory (CBRAM). The structure consists of an electrochemically active electrode (e.g. Ag or Cu) as the top electrode, an electrochemically inert electrode including W, TiN, doped-poly Si [13,30] as the bottom electrode. For the switching medium, an electrolyte material (e.g. chalcogenide materials) or conventional dielectric material (SiO_2, Al_2O_3 or a-Si) is used. When a positive voltage is applied to the electrochemically active electrode (Ag or Cu), the electrode can be oxidized. Here "oxidation" is defined broadly as the process of a metal atom losing an electron or electrons and forming a cation. The oxidized cations will migrate inside the switching medium in the direction towards the inert bottom electrode and eventually become reduced and deposited either inside the switching material or on top of the inert electrode. This leads to the growth of metal filaments that modulate the overall resistance of the device. Instead of cations, the other class of RRAM is based on anion migration, mainly oxygen ions, and are termed as VCM or simply oxide-RRAM. Here inert electrodes are

used, and binary transition metal oxides or complex perovskite oxides are used as the switching medium. The resistance modulation is achieved by the migration and relocation of oxygen ions, or equivalently, positively charged oxygen vacancies (V_Os). There are possibly two effects in general. In one case, the accumulation of the V_Os at the electrode/switching medium interface can change the Schottky barrier height hence resulting in a change of the resistance of the device. In the other case, the redistribution of V_Os can lead the formation and rupture of filaments consisting of V_O-rich regions. To date, high endurance up to 10^{12} [15], long retention of < 10 years, fast switching speed of <1ns [16], device size scaling down up 10 nm [17] has been demonstrated in oxide-based RRAM.

1.5.3 Memristor

The RRAM also falls in the category of memristor [31]. A mathematical framework of memristor (memristive systems) has been used to explain resistive switching effects in RRAM [30]. The resistance of memristor is determined by the instantaneous input (i) and one or a set of internal state variables (w) as shown in equation Eq. (1):

$$i = G(w, v)v \tag{1}$$

While a normal resistor (linear or non-linear, e.g. diode) may also be determined by an internal state variable (e.g. depletion width in a diode), the state of a normal resistor is determined directly by the instantaneous input (current or voltage). On the other hand, for a memristor, the input only determines the change rate of the state variable, rather than its overall value. The state variable equation of a memristor is shown in Eq. (2).

$$\dot{w} = f(w, v) \tag{2}$$

Eq. (2) essentially states that the device state is determined by a time integral of the input conditions, thus leading to a history-dependent resistance. Specifically, RRAM device operations such as ionic diffusion and drift during conduction path formation or rupture can be effectively modeled within the memristor framework, which not only explains the experimentally observed hysteresis effects but also provides a framework for fundamental understanding of the device physics and allows effective analytical and numerical models to be developed that can help predict device operation in a circuit and guide device optimization as shown in chapter 6 and chapter 7.

1.5.4 Research on RRAM

As mentioned in the previous section, RRAM has attracted significant interest among academia and industry due to its properties such as low switching voltage [13], non-volatility [16], high endurance [15], and multi-level characteristics [33-35]. In this thesis, we first focus on the fundamental understanding of RRAM operation through systematic experimental studies and multi-physics modeling, including detailed standard transport studies as well as non-standard methods such as noise analysis and retention failure studies. Following the mechanism analysis, we show enhanced performance of analog switching behavior through a doping process. Finally, we demonstrate potential application of RRAM devices in computing through a principal component analysis (PCA) network using RRAM arrays.

1.6 Organization of the Thesis

In chapter 1, we have discussed several topics related to current and alternative non-volatile memory technologies. In chapter 2, we focus on modeling of the dynamic resistive switching processes in RRAM. Specifically, by solving the three equations for oxygen vacancy transport, current continuity and Joule heating, we present a quantitative and accurate dynamic switching model that fully accounts for the resistive switching behaviors in RRAM in a unified framework.

In chapter 3, we perform systematic investigations on the resistance switching mechanism through detailed noise analysis, and show the resistance switching from high-resistance to low-resistance is accompanied by a semiconductor-to-metal transition mediated by the accumulation of oxygen-vacancies in the conduction path. From noise and transport analysis, we discuss the density of states and average distance of the V_Os at different resistance states, and develop a unified model to explain the conduction in both the HRS and the LRS and account for the resistance switching effects in these devices. Significantly, it is found that even though the conduction channel area is larger in the HRS, during resistive switching a localized region gains significantly higher V_O and dominates the conduction process.

In chapter 4, we report detailed retention studies of RRAM at high temperatures and the development of oxygen diffusion reliability model of oxide-RRAM devices. The device conductance in low resistance state (LRS) is constantly monitored at several temperatures (above 300℃). Specifically, the activation energy for oxygen vacancy diffusion can be directly calculated from the failure time versus temperature relationship. The experimental result is well explained by both analytical modeling and detailed numerical multi-physics simulation, which

confirm the filamentary nature of the conduction path in LRS. Finally, this experiment reveals the existence of multiple filaments in the same device.

In chapter 5, we show that doping tantalum oxide based RRAM with silicon atoms can facilitate oxygen vacancy formation and transport in the switching layer with adjustable ion hopping distance and drift velocity. The devices show larger dynamic ranges with easier access to the intermediate states while maintaining the extremely high cycling endurance (> 10^{10} set and reset), and are well suited for neuromorphic computing applications. We further provide a characterization methodology to quantitatively estimate the effective hopping distance of the oxygen vacancies.

In chapter 6, we investigate the feasibility of using RRAM devices to implement PCA network. First, the conductance changes of RRAM devices in a response of voltage pulses is studied and modelled with one internal state variable to trace the analog behavior of a RRAM. Secondly, we utilize Sanger's learning rule, which is derived from Hebb's learning rule, to a crossbar array of RRAM devices to perform PCA network and the weights distribution of the array is re-adjusted by the applied pulses calculated by the rule. We also examine the effect of device non-uniformity issue on the PCA network. In chapter 7, we discuss the experimental demonstration of unsupervised learning using RRAM networks and periphery circuitry based on the result of chapter 6.

In chapter 8, a brief discussion of this thesis and future works are mentioned.

1.7 **References**

[1] IC insights. (2015). Report Contents and Summaries. [Online] Avaiable: http://www.icinsights.com/services/mcclean-report/report-contents/. [2014, March 1].

[2] S. Nielson. (2014). Why oversupply could affect Micron's NAND flash products. [Online] Available: http://marketrealist.com/2014/04/micron-nand-flash-products/. [2014, March 1].

[3] C.-Y. Lu, K.-Y. Hsieh, and R. Liu, "Future Challenges of Flash Memory Technologies," *Microelectron. Eng.,* vol. 86, no. 3, pp. 283-86, Mar. 2009.

[4] A. Fazio, "Flash Memory Scaling," *MRS Bulletin*, vol. 29, no. 11, pp 814-17, Nov. 2004.

[5] K. Prall, "Scaling Non-Volatile Memory Below 30nm," *Non-Volatile Semicond. Mem. Work. 200*

22nd IEEE, pp. 5-10, Aug. 2007.

[6] K. Prall, and K. Parat, "25nm 64Gb MLC NAND Technology and Scaling Challenges," *Int. Elec. Dev. Meet. (IEDM), 2010 IEEE Inter.*, pp. 5.2.1–5.2.4, 2010.

[7] H.-S. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, and K. E. Goodson, "Phase Change Memory," *Proc. IEEE*, vol. 98, no. 12, pp. 2201–27, Dec. 2010.

[8] G.W. Burr, M. J. Breitwisch, M. Franceschini, D. Garetto, K. Gopalakrishnan, B. Jackson, B. Kurdi, C. Lam, L. A. Lastras, A. Padilla, B. Rajendran, S. Raoux, and R. S. Shenoy, "Phase Change Memory Technology," *J. Vac. Sci. Technol. B Microelec. Nanom. Struct.* vol. 28, no. 2, pp. 223–62, Mar. 2010.

[9] Y. J. Park, I. Bae, S. J. Kang, J. Chang, C. Park, "Control of Thin Ferroelectric Polymer Films for Non-Volatile Memory Applications," vol. 17, no. 4, pp. 1135–1163, 2010.

[10] Y. Huai and Y. Huai, "Spin-Transfer Torque MRAM (STT-MRAM): Challenges and Prospects," *AAPPS Bulletin*, vol. 18, no. 6, pp. 33–40, Dec. 2008.

[11] S. Tehrani, B. Engel, J. M. Slaughter, E. Chen, M. Deherrera, M. Durlam, P. Naji, R. Whig, J. Janesky, and J. Calder, “Recent Developments in Magnetic Tunnel Junction MRAM,” *IEEE Trans. Magn.*, vol. 36, no. 5, pp. 2752–57, Sep 2000.

[12] J. M. Slaughter, R. W. Dave, M. Deherrera, M. Durlam, B. N. Engel, J. Janesky, N. D. Rizzo, and S. Tehrani, “Fundamentals of MRAM Technology,” *J. Supercond: Incor. Nov. Magn.*, vol. 15, no. 1, Feb. 2002.

[13] S. H. Jo, K. Kim, and W. Lu, “High-Density Crossbar Arrays Based on a Si Memristive System,” *Nano lett.*, vol. 9, no. 2, pp. 870–74, Jan. 2009.

[14] M.-J. Lee, C. B. Lee, D. Lee, S. R. Lee, M. Chang, J. H. Hur, Y.-B. Kim, C.-J. Kim, D. H. Seo, S. Seo, U.-I. Chung, I.-K. Yoo, and K. Kim, “A fast, high-endurance and scalable non-volatile memory device made from asymmetric $Ta_2O_{(5-x)}$/$TaO_{(2-x)}$ bilayer structures.,” *Nat. Mater.*, vol. 10, no. 8, pp. 625–30, Aug. 2011.

[15] J. J. Yang, M.-X. Zhang, J. P. Strachan, F. Miao, M. D. Pickett, R. D. Kelley, G. Medeiros-Ribeiro, and R. S. Williams, “High switching endurance in TaO_x memristive devices,” *Appl. Phys. Lett.*, vol. 97, no. 23, p. 232102, Dec. 2010.

[16] A. C. Torrezan, J. P. Strachan, G. Medeiros-Ribeiro, and R. S. Williams, “Sub-nanosecond switching of a tantalum oxide memristor.,” *Nanotechnology*, vol. 22, no. 48, p. 485203, Dec. 2011.

[17] B. Govoreanu, G. S. Kar, Y. Chen, V. Paraschiv, S. Kubicek, A. Fantini, I. P. Radu, L. Goux, S. Clima, R. Degraeve, N. Jossart, O. Richard, T. Vandeweyer, K. Seo, P. Hendrickx, G. Pourtois, H. Bender, L. Altimime, D. J. Wouters, J. A. Kittl, M. Jurczak, B.- Leuven, and K. U. Leuven, “10x10nm^2 Hf/HfO_x Crossbar Resistive RAM with Excellent Performance , Reliability and Low-Energy Operation,” *Int. Elect. Dev. Meet. (IEDM)*, *2011 IEEE Inter.*, pp. 31.6.1–4, Dec. 2011.

[18] S. H. Jo, K.-H. Kim, and W. Lu, “Programmable resistance switching in nanoscale two-terminal devices.,” *Nano Lett*., vol. 9, no. 1, pp. 496–500, Jan. 2009.

[19] B. V. V Zhirnov, R. K. Cavin, L. F. Ieee, S. Menzel, E. Linn, S. Schmelzer, D. Bra, C. Schindler, and R. Waser, “Memory Devices : Energy – Space – Time Tradeoffs,” *Proc. IEEE*, vol. 98, no. 12, pp. 2185–2200, Oct. 2010.

[20] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, “The missing memristor found,” *Nature*, vol. 453, pp. 80–84, May 2008.

[21] S. Yu, H. Chen, B. Gao, J. Kang, and H. P. Wong, “HfO_x-Based Vertical Resistive Switching Random Access Memory Suitable for Bit-Cost-Effective Three-Dimensional Cross-Point Architecture,” *ACS Nano*,vol. 7, no. 3, pp. 2320–2325, 2013.

[22] K. Kim, S. Gaba, D. Wheeler, J. M. Cruz-albrecht, T. Hussain, N. Srinivasa, and W. Lu, “A Functional Hybrid Memristor Crossbar-Array/CMOS System for Data Storage and Neuromorphic Applications,” *Nano Lett.*, vol. 12, pp. 389-395, 2012.

[23] Q. Xia, W. Robinett, M. W. Cumbie, N. Banerjee, T. J. Cardinali, J. J. Yang, W. Wu, X. Li, W. M. Tong, D. B. Strukov, G. S. Snider, G. Medeiros-Ribeiro, and R. S. Williams, “Memristor - CMOS Hybrid Integrated Circuits for Reconfigurable Logic,” *Nano Lett.*, vol.9, no. 10, pp. 3640-3645, 2009.

[24] T. Chang, S. Jo, and W. Lu, “Short-Term Memory to Long-Term Memory Transition in a Nanoscale Memristor,” *ACS Nano*, vol. 5, no. 9, pp. 7669–7676, 2011.

[25] D. B. Strukov and K. K. Likharev, “CMOL FPGA: a reconfigurable architecture for hybrid digital circuits with two-terminal nanodevices,” *Nanotechnology*, vol. 16, no. 6, pp. 888–900, Jun. 2005.

[26] T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, J. K. Gimzewski, and M. Aono, “Short-term plasticity and long-term potentiation mimicked in single inorganic synapses.,” *Nat. Mater.*, vol. 10, no. 8, pp. 591–5, Aug. 2011.

[27] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, “Nanoscale Memristor Device as Synapse in Neuromorphic Systems,” *Nano Lett.*, vol. 10, pp. 1297–1301, 2010.

[28] F. Alibart, S. Pleutin, D. Guérin, C. Novembre, S. Lenfant, K. Lmimouni, C. Gamrat, and D. Vuillaume, "An Organic Nanoparticle Transistor Behaving as a Biological Spiking Synapse," *Adv. Funct. Mater.*, vol. 20, no. 2, pp. 330–337, Jan. 2010.

[29] R. Waser and M. Aono, "Nanoionics-based resistive switching memories.," *Nat. Mater.*, vol. 6, no. 11, pp. 833–40, Nov. 2007.

[30] S. H. Jo and W. Lu, "CMOS Compatible Nanoscale Nonvolatile Resistance Switching Memory," *Nano Lett.*, vol. 8, no. 2, pp. 392–397, 2008.

[31] L. O. Chua, "Memristor - The missing circuit element, " *IEEE Trans. Circuit Theory*, vol. CT-18, no. 5, pp. 507-519, 1971.

[32] L. O. Chua, and S.M. Kang, "Memristive Devices and Systems," *Proc. IEEE*, vol. 64, no. 2, Feb. 1976.

[33] K. Kim, S. H. Jo, S. Gaba, and W. Lu, "Nanoscale resistive memory with intrinsic diode characteristics and long endurance," *Appl. Phys. Lett.*, vol. 96, p. 053106, 2010.

[34] M. Wu, Y. Lin, W. Jang, C. Lin, and T. Tseng, "Low-Power and Highly Reliable Multilevel Operation in ZrO_2 1T1R RRAM," *IEEE Elec. Dev. Lett.*, vol. 32, no. 8, pp. 1026–1028, 2011.

[35] U. Russo, D. Kamalanathan, D. Ielmini, A. L. Lacaita, M. N. Kozicki, "Study of Multilevel Programming in Programmable Metallization Cell (PMC) Memory," IEEE Trans. Elec. Dev., vol. 56, no. 5, 2009.

Chapter 2.

Comprehensive Physical Model of Dynamic Resistive Switching in an Oxide Based RRAM

2.1 Introduction

As discussed in Chapter 1, RRAM devices are two-terminal electrical devices whose resistances can be changed through internal reconfigurations in the switching layer. Mathematically, these devices can be categorized in the memristor model where the resistive switching process can be described by the dynamic evolution of a set of internal state variables [1-3]. Such devices have been extensively studied for nonvolatile memory storage, neuromorphic computing, and implementation logic applications [4,5,6-9]. However, although a number of models have been proposed to describe the device behavior, they are either non-dynamic and can only predict steady-state properties [10] or oversimplified [3,11-15]. Providing an accurate, physics-based RRAM model that can explain and predict the rich dynamic resistive switching behaviors not only fills an urgent need that enables accurate simulation of large-scale RRAM systems but also can significantly improve our understanding of the different factors that drive the switching process and will be critical for continued optimization and design of this important class of devices.

In this chapter, we present a complete physical model that quantitatively and accurately describes the rich resistive switching behaviors in a tantalum-oxide-based bilayer RRAM. As memory devices, tantalum-oxide-based RRAM devices have shown excellent switching

performance between two discrete resistance levels, including extreme cycling endurances of over 10^{12} cycles and fast switching speeds below 10 ns [16,17]. By solving the dynamic transport equations of oxygen vacancies, we can precisely predict the resistive switching behaviors in tantalum-oxide-based RRAM devices in both DC and pulse operation modes using a single set of material-dependent parameters. More importantly, analog switching behaviors were also observed in the simulation and confirmed experimentally. Our quantitative analysis reveals that the SET process is driven by both the electric field and thermal effects, while RESET is mainly driven by thermal effects.

2.2 Resistive switching behavior of TaO_x/Ta_2O_5 bilayer

The tantalum-oxide-based bilayer RRAM consists of a highly resistive Ta_2O_5 layer on top of a less resistive TaO_x base layer sandwiched by top and bottom Pd electrodes (TE and BE) [18], as shown in Fig. 2.1(a). To explain the resistive switching behaviors, the concept of the formation/rupture of conductive filaments (CFs) has been generally accepted [19]. Here the filaments correspond to regions with high oxygen vacancy (V_O) concentration so the local electrical conduction increases significantly and can even become metallic [20]. The device can be set (from a low-conductance state to a high conductance state) or reset (from a high conductance to a low conductance) among different resistance states by controlling the properties (*e.g.* V_O concentration and shape) of the filament.

Our simulation starts from the state immediately after the electroforming process, where a continuous CF connects the TE and TaO_x layer in figure 2.1(c). The dynamic resistance switching processes are driven by V_O migration through three factors: local electric field, V_O

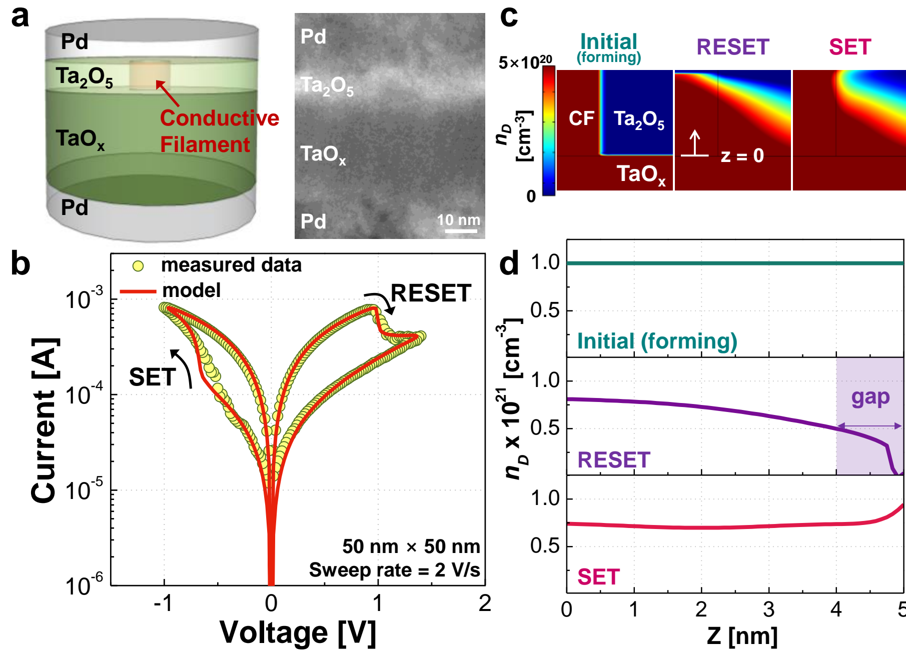


Figure 2.1. Modeling a tantalum oxide RRAM during set/reset. (a) Schematic and cross-sectional TEM images of the Pd/Ta_2O_5/TaO_x/Pd bilayer RRAM device. (b) Measured and calculated DC *I-V* characteristics of the Pd/Ta_2O_5/TaO_x/Pd device. The measured device size is 50 nm × 50 nm, and the voltage sweep speed is 2 V/s. (c) Calculated 2-D maps of n_D as well as (d) 1-D profiles of n_D along the center of the CF in the initial state, after reset, and after the set process. The depleted gap is determined as the position where $n_D = 5 \times 10^{20}$ cm^{-3}. The $z = 0$ position is the Ta_2O_5/TaO_x interface.

concentration gradient, and temperature gradient due to Joule heating [19-21]. Thus, the complete resistive switching process can be captured after self-consistently solving the three partial differential equations (PDEs) [15,21,22]: (1) a drift/diffusion continuity equation for V_O transport (equation (1)), (2) a current continuity equation for electrical conduction (equation (2)), and (3) a Fourier equation for Joule heating (equation (3)), as summarized in Figure 2. These three PDEs were self-consistently solved here through a numerical solver (COMSOL) to calculate the V_O concentration n_D, the electrostatic potential ψ, and the local temperature T. The details for the proposed model are discussed in Figures 2.2, 2.3 and 2.4.

Figure 2.1(b) shows the measured and calculated DC *I-V* characteristics during the set and reset processes. The reset transition starts near 0.9 V, and the resistance gradually increases,

finally achieving a state that is roughly one decade more resistive after reset. Similarly, the set transition occurs at a negative voltage, and both the calculated reset and set processes are accurately captured by the model. The physical nature of the set and reset processes can be studied by examining the n_D profiles, as shown in Figure 2.1(d). Specifically, during reset a gap of ~ 1 nm with a depleted V_O concentration was formed neat the TE, leading to the increase of the device resistance; while refilling the gap during set leads to the recovery of the high conductance.

2.3 **Details of the model**

To describe the drift/diffusion migration of V_O, the model proposed by Mott and Gurney was employed [23]. The diffusion coefficient is given by $D = 1/2 \cdot a^2 \cdot f \cdot \exp(-E_a/kT)$, and the drift velocity is given by $v = a \cdot f \cdot \exp(-E_a/kT) \cdot \sinh(qaE/kT)$, where f is the escape-attempt frequency (10^{12} Hz) [23], a is the effective hopping distance (0.1 nm), and E_a is the activation energy for migration.

Considering drift and diffusion, the time-dependent evolution of the V_O concentration (n_D) can be expressed by the following continuity equation:

$$\frac{\partial n_D}{\partial t} = \nabla \cdot (D \nabla n_D - v n_D + D S n_D \nabla T) \, . \qquad (1)$$

• Dependent variables		**• Constants**	
n_D	Concentration of V_o [cm^{-3}]	a	Hopping distance, 0.1 nm
T	Temperature [K]	f	Escape-attempt frequency, 10^{12} Hz
ψ	Potential [V]	E_a	Diffusion barrier, 0.85 eV

• Oxygen vacancy transport

Eq.(1) $$\frac{\partial n_D}{\partial t} = \nabla \cdot (D\nabla n_D - v n_D + DSn_D \nabla T)$$

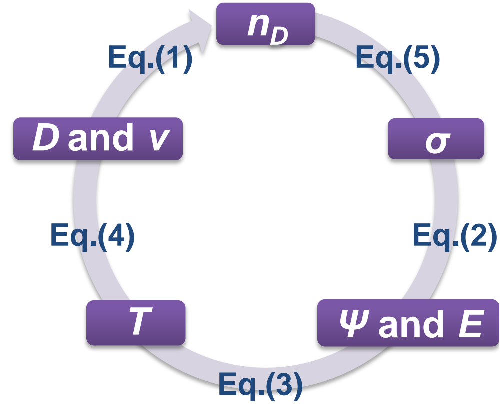
• Current continuity

Eq.(2) $$\nabla \cdot \sigma \nabla \psi = 0$$

• Heat (Joule heating)

Eq.(3) $$-\nabla \cdot k_{th} \nabla T = J \cdot E = \gamma \cdot \sigma |\nabla \psi|^2$$

(γ = 1 for DC, and γ = 2 for AC simulation)



• Parameters - Eqs.(4)

$D = 1/2 \cdot a^2 \cdot f \cdot \exp(-E_a / kT)$	Diffusivity of V_o [cm^2s^{-1}]
$v = a \cdot f \cdot \exp(-E_a / kT) \cdot \sinh(qaE / kT)$	Drift velocity of V_o [cm/s]
$S = -E_a / kT^2$	Soret diffusion coefficient [1/K]

Figure 2.2. Equations and parameters in the proposed model. Three PDEs are self-consistently solved with a numerical solver.

Here, $D\nabla n_D$ and vn_D are terms for Fick diffusion flux and drift flux, respectively. The $DSn_D\nabla T$ term is the Soret diffusion flux, where S is the Soret coefficient. Soret diffusion, also referred to as thermophoresis, is the movement of molecules along a temperature gradient and is commonly observed in liquids or molecular solutions [24]. However, its role in solid oxides has recently been emphasized [25,26]. The Soret diffusion term describes the tendency for V_O to move toward the hotter region in a temperature gradient [25]. It is noted that the Soret diffusion term has positive sign because oxygen move towards low temperature region so oxygen vacancies, its counterpart, move towards high temperature. We found the Soret diffusion term needs to be included in the transport continuity equation to achieve accurate simulation. Equation

(1) can be solved when coupled with the current continuity equation for electrical conduction, such that

$$\nabla \cdot \sigma \nabla \psi = 0 \tag{2}$$

where σ is the electrical conductivity; and with the steady-state Fourier equation for Joule heating,

$$-\nabla \cdot k_{th} \nabla T = J \cdot E = \gamma \cdot \sigma \,|\, \nabla \psi \,|^2 \tag{3}$$

where k_{th} is the thermal conductivity. The transient term of the Fourier equation ($\rho Cp \bullet \partial T/\partial t$) was disregarded in equation (3) because the values of the specific heat capacity (C_p) and mass density (ρ) for the different n_D values are not known. Instead, an additional fitting parameter γ is introduced to describe the transient effect which will be different for DC and AC programming conditions, where $\gamma = 1$ and $\gamma = 2$ were used for DC and AC, respectively, in the simulation.

The three PDEs in Eqs. (1)–(3) were self-consistently solved by a numerical solver (COMSOL) to calculate n_D, ψ (potential), and T (temperature). To solve equations (2) and (3), models for the electrical conductivity (σ) and thermal conductivity (k_{th}) are required. To this end, both σ and k_{th} are assumed to depend on n_D, as shown in Figure 2.3 [11,27]. The electrical conductivity is given by the Arrhenius equation [28], $\sigma = \sigma_o \cdot \exp(-E_{AC}/kT)$, where σ_o is a pre-exponential factor and E_{AC} is the activation energy for conduction. As shown in Figure 3a, σ_o is assumed to linearly increase from 10 to 940 Ω^{-1} cm^{-1} with increasing n_D. In addition, Figure 2.2(b) shows the conduction activation energy E_{AC} used in the calculations. The activation energy is -0.006 eV for high n_D and linearly increases to 0.05 eV with decreasing n_D; these values correspond to the measured values at the LRS and the HRS, respectively. Moreover, a

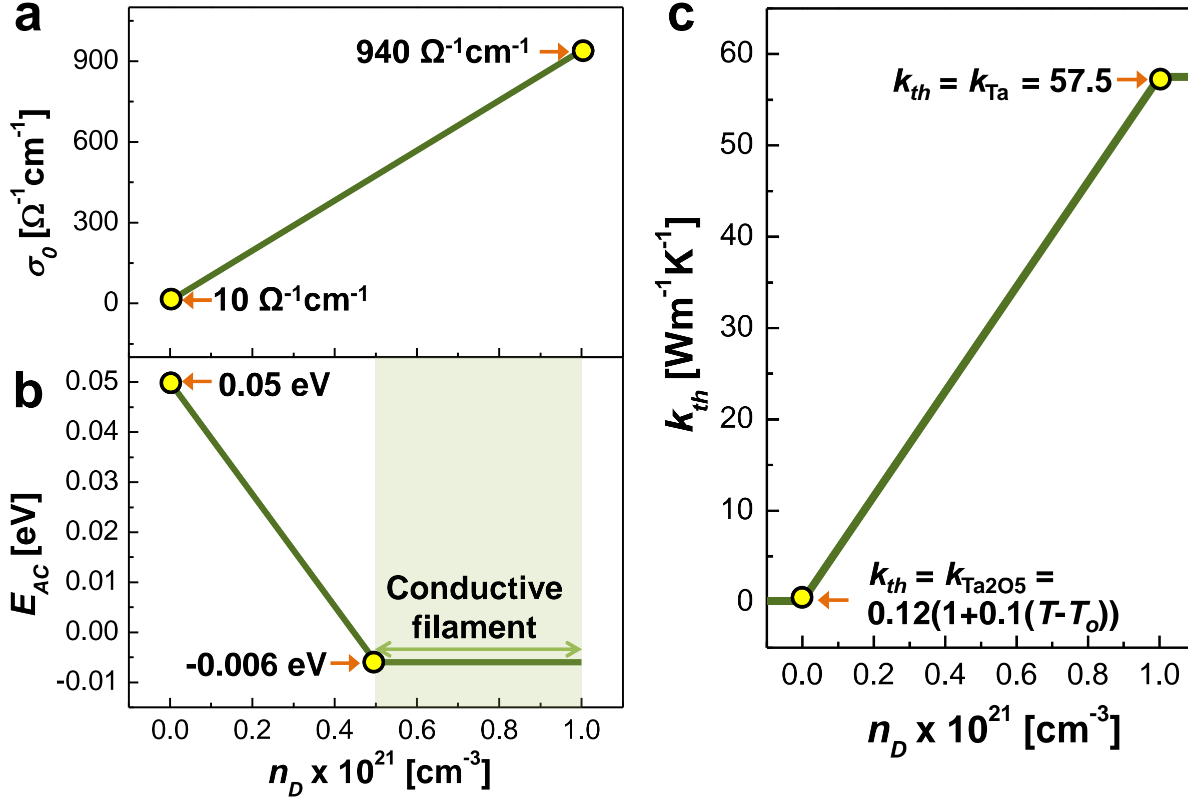


Figure 2.3. Parameters from measurements and assumptions. (a) Electrical conductivity pre-exponential factor σ_0, (b) assumed activation energy for conduction E_{AC}, and (c) assumed thermal conductivity k_{th} as a function of local V_O density n_D.

linear dependence of k_{th} on n_D is assumed on the basis of the Wiedemann-Franz law, as shown in Figure 2.2(c) [27]. The minimum value for n_D = 0 refers to the thermal conductivity of the insulating Ta_2O_5, k_{Ta2O5} = 0.12 W m^{-1} K^{-1} for T_0 = 300 K [29]. In addition, linear temperature dependence of k_{Ta2O5} is assumed, as $k_{Ta2O5} = 0.12\bullet(1+\lambda(T-T0))$, where λ = 0.1 is the linear thermal coefficient. The maximum k_{th} value at high n_D corresponds to that of the metallic CF, i.e., the thermal conductivity of tantalum k_{Ta} = 57.5 W m^{-1} K^{-1}. Here, an approximate maximum V_O density of 1×10^{21} cm^{-3} is chosen because the V_O concentration of the metallic Magnelli phase

Ti_4O_7 is on the order of 10^{21} cm^{-3} [30]. Although the particular choice of the maximum V_O value and the linear approximations of σ_0, E_{AC}, and k_{th} appear to be over-simplified, the calculation results based on these assumptions show good consistency with the experimental data. Therefore, these assumptions do not limit the validity of the calculations.

In the actual calculations, the axisymmetric geometry of the device allowed the 3-D

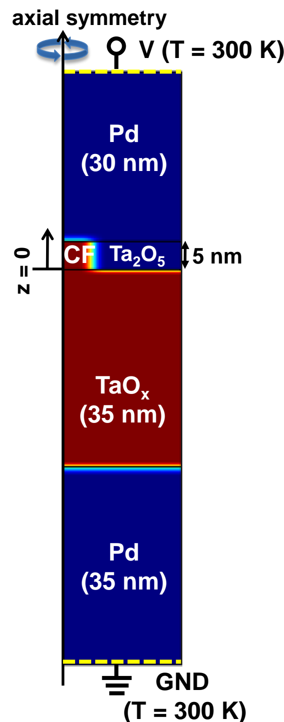


Figure 2.4. Simulated geometry used in the calculation. The axisymmetric geometry reduces the problem from 3-D to 2-D. A uniform doping concentration of $n_D = 1 \times 10^{21}$ cm^{-3} was assumed within the CF and the TaO_x layer as the initial state right after electroforming.

problem to be reduced to a 2-D solution with a radial coordinate and a vertical coordinate, as shown in Figure 2.4. The oxide bilayer materials (Ta_2O_5 and TaO_x layers) are sandwiched by two electrodes, and all layers including the two electrodes are considered in the calculations (for the

Pd TE and BE, $\sigma_{Pd} = 9.5 \times 10^4\ \Omega^{-1}\ cm^{-1}$ and $k_{Pd} = 71.8\ W\ m^{-1}\ K^{-1}$). The boundary conditions for equation (2) are $\psi = 0$ and $\psi = V$ at the BE and TE, respectively. The outermost boundaries of the two electrodes are defined with boundary conditions T = 300 K; this assumption is reasonable because the electrode area is generally large with respect to the CF and can be sufficiently cooled. For the V_O drift/diffusion, no V_O flux was assumed at the TE/Ta_2O_5 and TaO_x/BE interfaces. A uniform concentration of $n_D = 1 \times 10^{21}\ cm^{-3}$ was defined within the CF and TaO_x layer as the initial state (i.e., the state immediately after forming). The CF size was set to a diameter of 5 nm, in agreement with direct evaluations using conductive atomic force microscopy [31].

2.4 **Conclusion**

By solving the local electric field, temperature and V_O concentration self-consistently, we developed a complete and accurate physics-based model that quantitatively explains the dynamic resistive switching process. Significantly, the model reveals that the conducting filament is ruptured and formed locally inside the switching layer, and the set process involves field-driven filament formation followed by filament expansion, while reset process is dominated by thermal-driven filament rupture followed by gap widening. The competition between the drift and diffusion components during reset can lead to different resistive switching characteristics. The proposed model allows accurate prediction of resistive switching characteristics for both DC and AC input signals, and was able to reproduce the analog switching behavior. We believe such in-depth analysis of the resistive switching process not only provides a reliable and accurate physical picture of the resistive switching process but also produces much-needed guidelines for

continued design and optimization of this important class of devices for memory and logic applications.

2.5 **References**

[1] Chua, L. O. Memristor - The missing circuit element. IEEE Trans. Circuit Theory 1971, CT-18, 507-519.

[2] Di Ventra,; M. Pershin, Y. V.; Chua, L. O. Circuit elements with memory: memristors, memcapacitors, and meminductors. Proceeding of the IEEE 2009, 97, 1717-1724.

[3] Chua, L. Resistance switching memories are memristors. Appl. Phys. A 2011, 102, 765-783.

[4] Jo, S. H.; Chang, T.; Ebong, I.; Bhadviya, B. B.; Mazumder, P.; Lu, W. Nanoscale memristor device as synapse in neuromorphic systems. Nano Lett. 2010, 10, 1297-1301.

[5] Yu, S.; Wu, Y.; Jeyasingh, R.; Kuzum, D.; Wong, H. –S. P. An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation. IEEE Trans. Elect. Dev. 2011, 58, 2729-2737.

[6] Ambrogio, S.; Balatti, S.; Nardi, F.; Facchinetti, S.; Ielmini, D. Spike-timing dependent plasticity in a transistor-selected resistive switching memory. Nanotechnology 2013, 24 384012-384012.

[7] Chang, T.; Jo, S. –H.; Lu, W. Short-term memory to long-term memory transition in a nanoscale memristor. ACS Nano 2011, 5, 7669-7676.

[8] Alibart, F.; Zamanidoost, E.; Strukov, D. B. Pattern classification by memristive crossbar circuits using ex situ and in situ training. Nature Communications 2013, 4, 2072.

[9] Zamarreno-Ramos, C.; Camunas-Mesa L. A.; Perez-Carrasco, J. A.; Masquelier T.; Serrano-Gotarredona, T.; Linares-Barranco B. On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex. Front Neurosci. 2011, 5, 26.

[10] Strukov, D. B.; Williams, R. S. Exponential ionic drift: fast switching and low volatility of thin-film memristors. Appl. Phys. A 2009, 94, 515.

[11] Ielmini, D. the universal set/reset characteristics of bipolar RRAM by field- and temperature-driven filament growth. IEEE Trans. Electron Dev. 2011, 58, 4309-4317.

[12] Chang, T.; Jo, S. –H.; Kim, K. –H.; Sheridan, P.; Gaba, S.; Lu, W. Synaptic behaviors and modeling of a metal oxide memristive device. Appl. Phys. A 2011, 102, 857-863.

[13] Eshraghian K.; Kavehei, O.; Cho, K. –R.; Chappell, J. M.; Iqbal, A.; Al-Sarawi, S. F.; Abbott D. Memristive device fundamental and modeling: application to circuits and systems simulation. Proceeding of the IEEE 2012, 100, 1991-2007.

[14] Yu, S.; Wong, H.–S. P. A phenomenological model for the reset mechanism of metal oxide RRAM. IEEE Electron Device Lett. 2010, 31, 1455-1457.

[15] Nardi, F.; Balatti, S.; Larentis, S.; Ielmini, D. Complementary switching in metal oxides: Toward diode-less crossbar RRAMs. 2011 IEEE International Electron Devices Meeting (IEDM), Washington, DC, Dec. 5-7, 2011; 31.1.1-31.1.4.

[16] Lee, M. -J.; Lee, C. B.; Lee, D.; Lee, S. R.; Chang, M.; Hur, J. H.; Kim, Y. –B.; Kim, C. –J.; Seo, D. H.; Seo, S.; Chung, U. –I.; Yoo, I. –K.; Kim, K. A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O5-x/TaO2-x bilayer structures. Nat. Mater. 2011, 10, 625-630.

[17] Torrezan, A. C.; Strachan, J. P.; Medeiros-Ribeiro, G.; Williams, R. S. Sub-nanosecond switching of a tantalum oxide memristor. Nanotechnology, 2011, 22, 485203.

[18] Yang, Y.; Choi, S.; Lu, W. Oxide heterostructure resistive memory. Nano Lett. 2013, 13, 2908-2915.

[19] Waser, R.; Dittmann, R.; Staikov, G.; Szot, K. Redox-based resistive switching memories-Nanoionic mechanisms, prospects, and challenges. Adv. Mater. 2009, 21, 2632-2663.

[20] Larentis, S.; Cagli, C.; Nardi, F.; Ielmini, D. Microelectron. Eng. 2011, 7, 1119-1123.

[21] Larentis, S.; Nardi, F.; Balatti, S.; Gilmer, D. C.; Ielmini, D. Resistive switching by voltage-driven ion migration in bipolar RRAM – Part II: Modeling. IEEE Trans. Electron Dev. 2012, 59, 2468-2475.

[22] Kim, S.; Kim, S. –J.; Kim, K. M.; Lee, S. R.; Chang, M.; Cho, E.; Kim, Y. –B.; Kim, C. J.; Chung, U. –I.; Yoo, I. –K. Physical electro-thermal model of resistive switching in bi-layered resistance-change memory. Scientific Reports 2013, 3, 1680.

[23] Mott, M. F.; Gurney, R. W. *Electronic Processes in Ionic Crystals.* Dover: U.K. 1948.

[24] Duhr, S.; Braun, D. Why molecules move along a temperature gradient. *Proc. Natl. Acad, Sci. U.S.A.* 2006, *103*, 19678-19682.

[25] Strukov, D. B.; Alibart, F.; Williams, R. S. Thermophoresis/diffusion as a plausible mechanism for unipolar resistive switching in metal-oxide-metal memristors. *Appl. Phys. A* 2012, *107*, 509-518.

[26] Mickel, P. R.; Lohn, A. J.; Choi, B. J.; Yang, J. J.; Zhang, M. X.; Marinella, M. J.; James, C. D.; Williams, R. S. A Physical Model of Switching Dynamics in Tantalum Oxide Memristive Devices. *Appl. Phys. Lett.* 2013, *102*, 223502.

[27] Larentis, S.; Nardi, F.; Balatti, S.; Gilmer, D. C.; Ielmini, D. Resistive Switching by Voltage-Driven Ion Migration in Bipolar RRAM-Part II: Modelling. *IEEE Trans. Electron Devices* 2012, *59*, 2468-2475.

[28] Ielmini, D.; Nardi, F.; Cagli, C. Physical Models of Size-Dependent Nanofilament Formation and Rupture in NiO Resistive Switching Memories. *Nanotechnology* **2011**, *22*, 254022.

[29] Henager, C. H.; Pawlewicz, W. T. Thermal Conductivities of Thin, Sputtered Optical Films. *Applied Optics* **1993**, *32*, 91-101.

[30] Kwon, D. –H.; Kim, K. M.; Jang, J. H.; Jeon, J. M.; Lee, M. H.; Kim, G. H.; Li, X. –S.; Park, G. –S.; Lee, B.; Han, S.; Kim, M.; Hwang, C. S. Atomic Structure of Conducting Nanofilaments in TiO_2 Resistive Switching Memory. *Nat. Nanotechnol.***2010**, *5*, 148-153.

[31] Yun, J.–B.; Kim, S.; Seo, S.; Lee, M. –J.; Kim, D. –C.; Ahn, S. –E.; Park, Y.; Kim, J.; Shin, H. Random and Localized Resistive Siwtching Observation in Pt/NiO/Pt. *Phys. Stat. Sol. (RRL)* **2007**, *1*, 280-282.

Chapter 3.

Random Telegraph Noise and Resistance Switching Analysis of Oxide Based RRAM

3.1 Introduction

Even though RRAM has been widely viewed as a promising candidate for future data storage applications [1-4], important questions regarding the nature of the conduction channels and the switching dynamics still remain under debate. In the previous chapter, we discussed a complete physical model that quantitatively describes the rich resistive switching behaviors in a tantalum-oxide-based bilayer RRAM. However, direct observation of the conduction channels formed by oxygen vacancy (V_O) redistribution is challenging, and electrical characterizations only provide limited information. In this chapter, we perform systematic investigation of the resistance switching mechanism in a TaO_x based RRAM through detailed noise analysis, and show the resistance switching from high-resistance to low-resistance is accompanied by a semiconductor-to-metal transition mediated by the accumulation of oxygen-vacancies in the conduction path. Specifically, pronounced random-telegraph noise (RTN) with values up to 25% was observed in the device high-resistance state (HRS) but not in the low-resistance state (LRS). Through time-domain and temperature dependent analysis, we show the RTN effect shares the same origin as the resistive switching effects, and both can be traced to the (re)distribution of oxygen vacancies (V_Os). From noise and transport analysis we further obtained the density of

states and average distance of the V_Os at different resistance states, and developed a unified model to explain the conduction in both the HRS and the LRS and account for the resistance switching effects in these devices. Significantly, it was found that even though the conduction channel area is larger in the HRS, during resistive switching a localized region gains significantly higher V_O and dominates the conduction process. These findings reveal the complex dynamics involved during resistive switching and will help guide continued optimization in the design and operation of this important emerging device class.

3.2 Device Fabrication and Measurement Setup

The resistive memory devices studied here is based on a $Pd/TaO_x/Ta_2O_5/Pd$ structure, in which the Ta_2O_5 layer acts as the switching layer and the TaO_x layer acts as the base layer that controls the device on-state resistance and provide needed oxygen vacancies for resistive switching [1,5]. The devices were fabricated on a Si/SiO_2 substrate with a 100 nm thermal SiO_2 layer. The bottom electrode consisting of 5nm-thick NiCr and 35nm-thick Pd was first patterned by e-beam lithography and deposited by e-beam evaporation. The TaO_X base layer (~50 nm) was deposited by direct current (DC) reactive sputtering of a Ta metal target with Ar/O_2 (3% oxygen partial pressure) gas mixture at 400°C, followed by the Ta_2O_5 switching layer (~5 nm) deposition by radio frequency (RF) sputtering of a Ta_2O_5 ceramic target at room temperature. Finally, the top-electrode consisting of 30 nm thick Pd and 20nm thick Au was patterned by e-beam lithography and deposited by the e-beam evaporation, forming a crossbar structure with the bottom electrode. Devices with different sizes from 50 nm × 50 nm to 5 μm × 5 μm were

fabricated and tested, as shown in the inset of Fig 3.1(a). During testing, the bias voltage was applied to the top electrode with the bottom electrode grounded. Reliable bipolar resistance switching characteristics with current – voltage (I-V) curves as shown in Fig. 3.1(a) were obtained after an electroforming process with up to -6 V. The electrical characterizations and noise measurements were performed using a custom-built measurement system and a temperature-variable probe station (Desert Cryogenics TTP4).

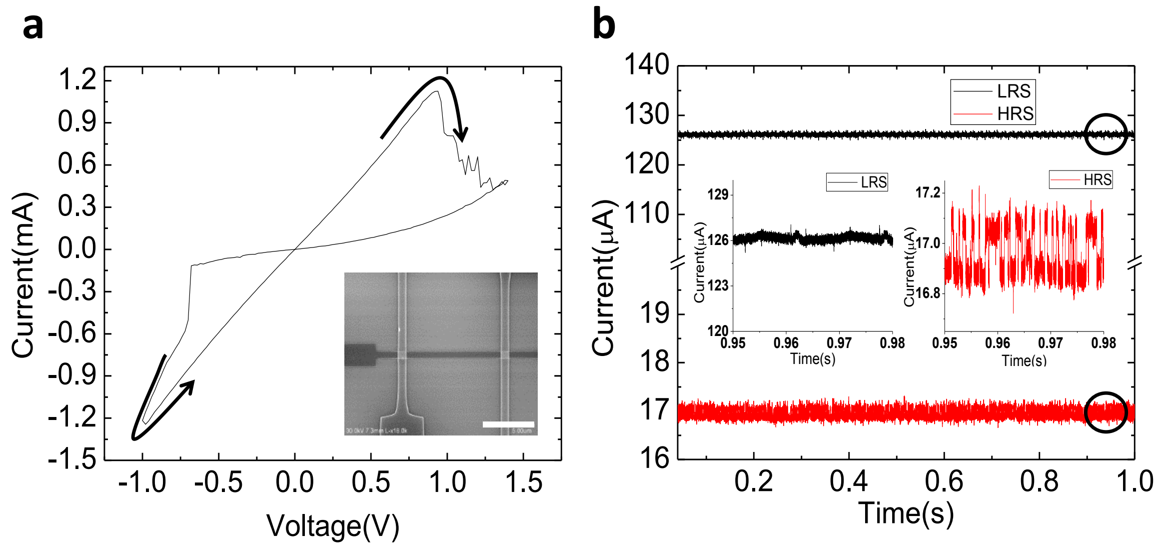


Figure 3.1. (a) I-V characteristics of a typical device showing the bipolar switching effects. Inset: SEM image of the device. Scale bar is 5μm. (b) Current-time plots measured at 0.1 V for LRS and HRS, respectively. Insets: zoomed-in plots of the circled areas for LRS (left) and HRS (right), showing pronounced RTN in HRS.

3.3 Statistical and Temperature Dependent Studies of RTN

A typical device switches to the LRS at around -0.7 V and switches back to the HRS at around 1.2 V (Fig. 3.1(a)). Noise measurements were performed by monitoring the device

current under a low, constant bias voltage of 0.1 V (Fig. 3.1(b)). Significantly, in the HRS, abrupt current jumps between two (or more) discrete current levels can be regularly observed (Fig.

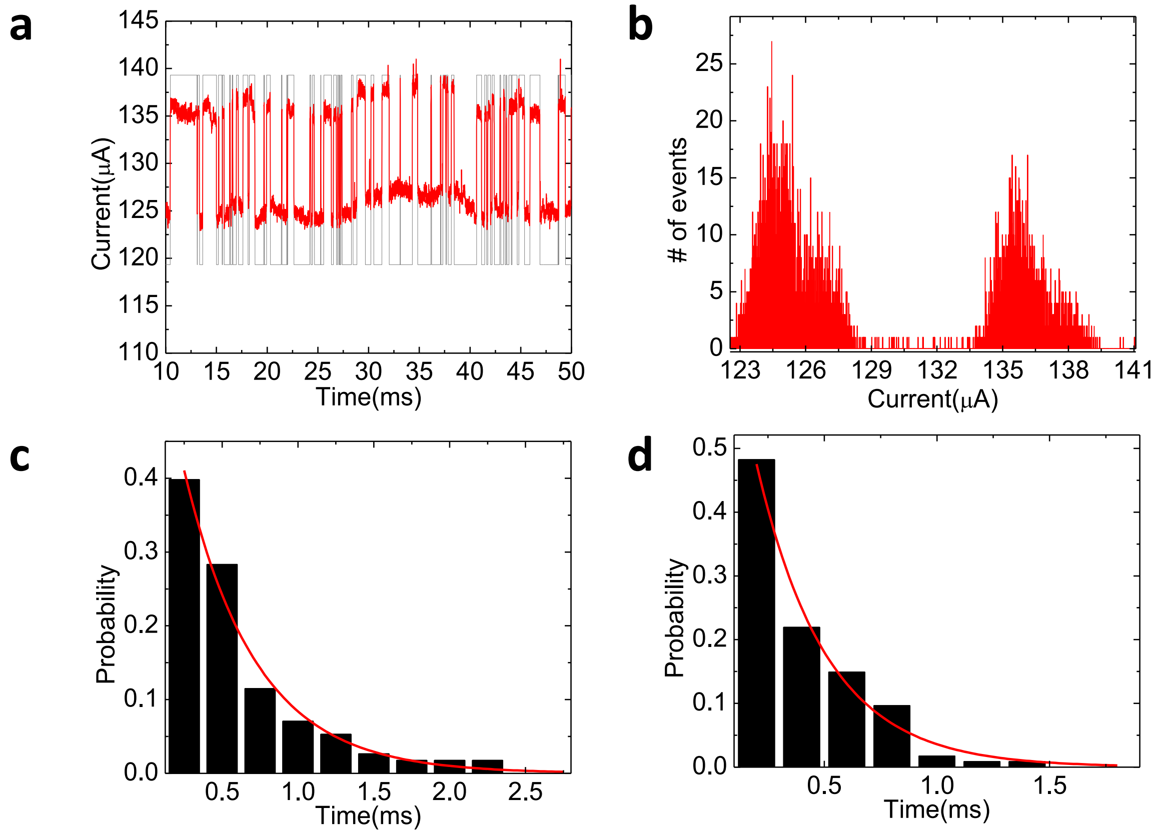


Figure 3.2. (a) Time-domain analysis of the RTN behavior showing raw data (red) and reproduced data (grey) based on the capture program. (b) Histograms of current vs. occurrence showing a bimodal distribution corresponding to the two current levels causing RTN. (c), (d) Histograms of the dwell times in the upper (c) and lower (d) current levels. The red lines are Poisson fits using as the only fitting parameter.

3.1(b), right inset), corresponding to significant RTN. On the other hand, RTN was not observed in the LRS despite the fact that the resistance values between the LRS and HRS differ by only ~10. Out of more than 70 devices tested, all devices showed similar behaviors.

Statistical and temperature dependent studies were carried out to reveal the nature of the observed RTN behaviors in HRS. The RTN data recorded in time-domain from a 500nm x

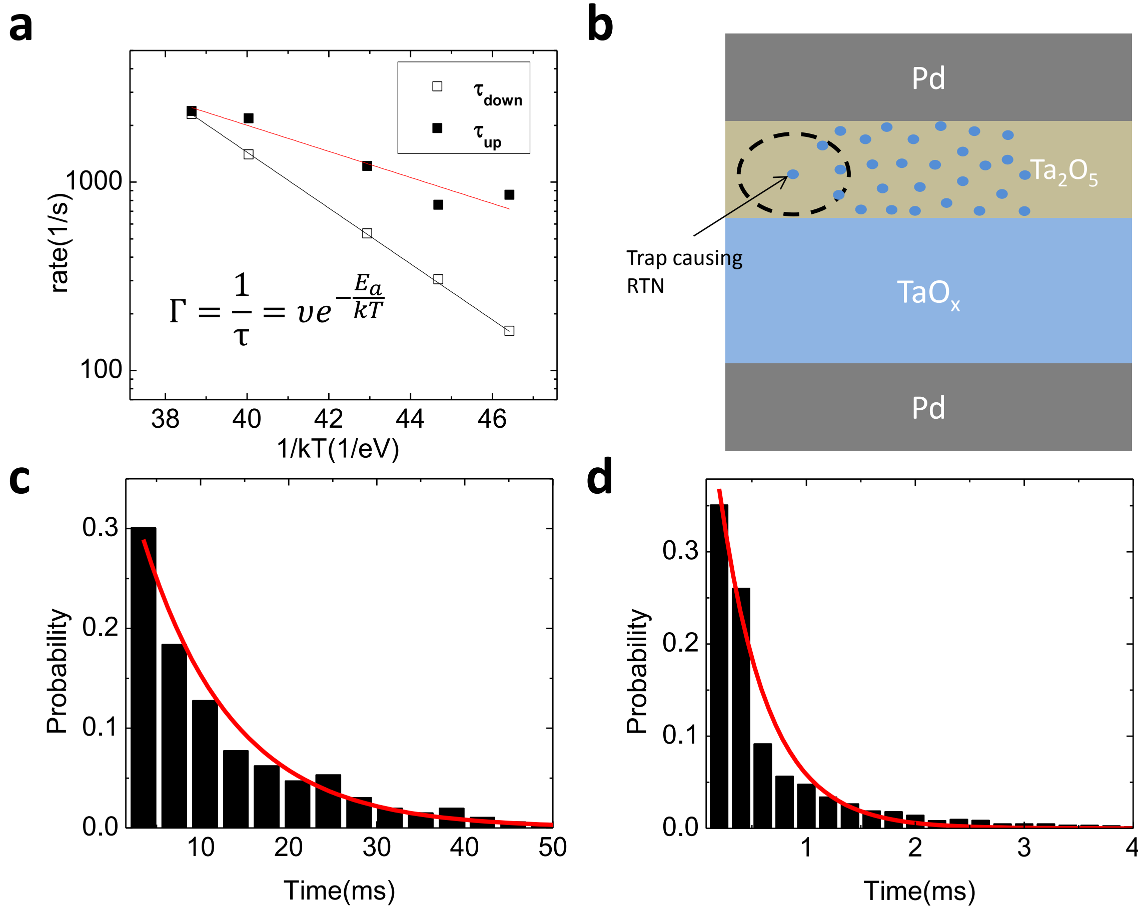


Figure 3.3. **(a) Temperature dependence of the characteristic dwell times in the upper and lower current levels. The lines are fits following the Arrhenius equation. (b) Schematic of the cause for RTN. The trapping and detrapping of a trap site near the channel leads to jumps in discrete current levels. The dashed circle represents the area that may be electrostatically depleted by the trapped electron. (c),(d) Histograms and fits of the dwell times at the upper current levels at 250 K and 300 K, respectively.**

500nm cell in HRS is shown in Fig. 3.2(a) (red line). Current jumps of ~10% between two discrete states are clearly observed. Plotting the current readings vs. occurrence clearly reveals a bimodal resistance distribution corresponding to two metastable states, as shown in Fig. 3.2(b). RTN as large as 25% has been observed in these devices in HRS. Such high noise levels can have a significant impact on device operation and deserve careful analysis. To analyze the RTN signal, a custom MATLAB code was used to capture the current jumps and measure the dwell time at each state [6]. The algorithm tests two hypotheses (whether a switching event happens or

not) within a given time window and maximizes the probability of both hypotheses in the presence of white noise. Using the maximum likelihood estimates, one can judge if a switching event occurs or not by examining the difference between the two maximized likelihood values and comparing it to a pre-set threshold value. If a switching event is judged to have occurred, the switching time is also recorded as the time that leads to the maximum likelihood. The reliability of this algorithm is verified in Fig. 3.2(a), where every current jump was successfully captured by the code (gray line) with no false positive or negative alarms. The time the device spent in each state is then recorded and analyzed, the histograms of which are plotted in Fig. 3.2(c), (d). For both metastable current states, the dwell time distribution can be fitted well with a Poisson distribution (solid lines). Following standard statistics, the Poisson distribution describes stochastic events such that the probability that an event (current jump) occurs within Δt at a given time t is [7, 8]

$$P(t) = \frac{\Delta t}{\tau} e^{-t/\tau} \tag{1}$$

where τ is the characteristic time constant. Results in Fig. 3.2(c), (d) can be fitted with Eq. (1) using τ as the only fitting parameter, showing the RTN noise can be well explained by stochastic events. This observation is consistent with the hypothesis that the RTN in RRAM is caused by electron trapping and detrapping at a trap site near the conduction path [9], as has been used to explain RTN in aggressively scaled MOSFET transistors [10] and nanowire devices [11].

Following the discussions in Ref. [9], the RTN in the RRAM can be explained as follows: when an electron falls into a trap near the conduction path in the RRAM device, it depletes the conduction path and causes the current to change to a lower level, as illustrated in Fig. 3.3(b). Similarly, the current will change to the higher level when the trapped electron is

detrapped. Since the trapping/detrapping events are thermally activated and stochastic in nature, these processes lead to the stochastic jumps between the current levels and the observed Poisson distribution of dwell times with well-defined characteristic time constants, if the current is affected by one dominate trap.

More evidence supporting this theory can be obtained from temperature dependent studies. Briefly, time-domain RTN data were analyzed at different temperatures and the extracted characteristic dwell time, τ, was then plotted as a function of temperature (T), as shown in Fig. 3.3(a). Typical results obtained at 250 K and 300 K are shown in Fig. 3.3(c),(d). Both τ_{up} and τ_{down}, corresponding respectively to the characteristic time constants in the upper current (detrapped) level and lower current (trapped) level, were recorded and analyzed. It can be seen that both time constants increase as the temperature is decreased [8, 12, 13]. Fitting the τ-T curves using an Arrhenius-type relationship produces the activation energies for the electron trapping and detrapping processes, which are 0.16 eV for electron trapping and 0.34 eV for electron detrapping for the device in Fig. 3.3. The schematic of the trapping/detrapping processes that lead to RTN is shown in Fig. 3.3(b).

3.4 Electron Transport Experiment in LRS and HRS.

It is interesting to note that RTN in general is not observed in LRS, even though the resistance values between LRS and HRS do not seem to differ significantly. We believe the different noise behaviors in HRS and LRS can shed light into the mechanism behind resistive switching in these RRAM devices, and the evolution of the noise characteristics correlate with the evolution of the conduction channels. In metal oxide based devices the conduction is

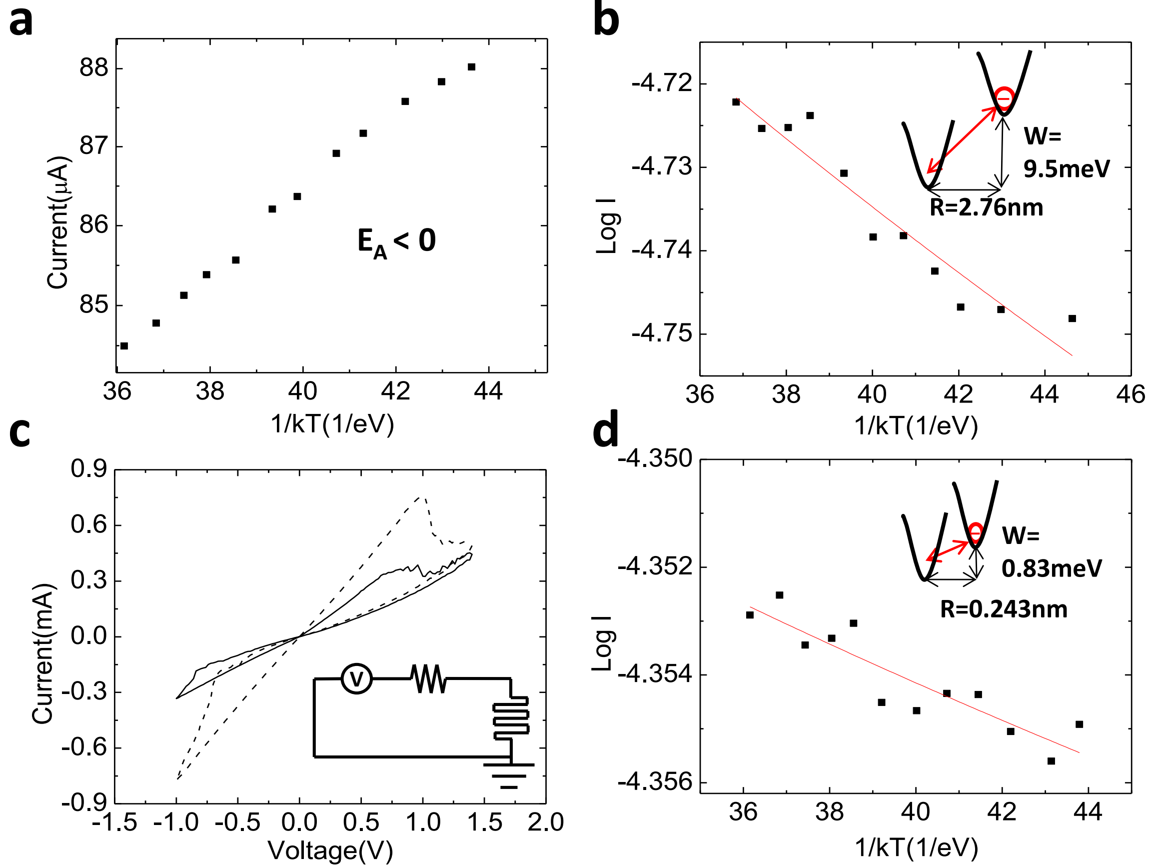


Figure 3.4. (a) Temperature dependence of electron transport in LRS. (b) Temperature dependence of electron transport in HRS. Inset: Schematic of the hopping process. (c) Solid line: I-V characteristics without the series resistor, showing switching between HRS and LRS; dashed line: I-V characteristics with a 1 kΩ series resistor. The device is programmed to an intermediate state instead. Inset: The circuit schematic. (d) Temperature dependence of electron transport in the intermediate state. Inset: Hopping with more closely spaced trap sites and lower hopping energy in the intermediate state compared to the HRS.

believed to be through conduction channels formed in regions with higher concentration of V_Os. Changes in V_O distribution lead to changes in resistance states, with the LRS state having a higher V_O concentration in the conduction channels compared with HRS [1, 2, 5, 14]. Consequently, these changes in trap distributions can also lead to different noise characteristics. To verify this hypothesis, temperature dependent studies of conduction through the LRS and HRS states were carried out, as shown in Fig. 3.4(a), (b). In the LRS, the current decreases as the

temperature is increased, suggesting the conduction channel is metallic with a negative activation energy for electron transport. On the other hand, in the HRS the device current increases as the temperature is increased, suggesting the conduction channel is semiconducting with a positive activation energy for electron transport.

In the HRS, the electron transport is believed to be facilitated by electron hopping mediated through the V_O trap sites [15], and the conduction can be explained using the variable-range hopping (VRH) model. If we define 1/α as the decay length of the electron wave function and R as average hopping distance, then the electron can hop to another trap site within the decay length if αR is equal to or less than unity, as shown in the inset of Fig. 3.4(b). During the process, the electron overcomes the hopping energy barrier, *W*, which corresponds to the energy difference between the two trap sites, assisted by thermal energy. As a result, the conductivity can be written at low electrical field as [16]

$$\sigma = 2e^2R^2N(E_F)\upsilon_{ph}\exp(-2\alpha R - \frac{W}{k_BT}) \quad (2)$$

where υ_{ph} is a factor related to electron-phonon interaction.. Following Mott's approach, *W* is in turn related to the density-of-states of the traps $N(E_F)$ through [16]

$$W = \frac{3}{4\pi R^3 N(E_F)} \quad . \quad (3)$$

Plugging Eq. (3) in to Eq. (2), and the most probable conduction occurs when

$$R = \left[\frac{9}{8\pi\alpha N(E_F)kT}\right]^{1/4}. \quad (4)$$

At this condition, the conductivity can be written as

$$\sigma = 2e^2R^2N(E_F)\upsilon_{ph}\exp\left(-\frac{T_o}{T}\right)^{1/4} \quad , \quad (5)$$

where $T_o = \frac{18\alpha^3}{k_B N(E_F)}$. We note Eq. (5) is the well-known Mott equation for transport through disordered systems in three-dimension.

The temperature dependent results in Fig. 3.4(b) can be well-fitted with Eq. (5), which leads to an extracted density of localized (oxygen vacancy) trap states $N(E_F)$ of 1.2×10^{21} $eV^{-1}cm^{-3}$, and average hopping distance R of 2.8 nm and activation energy at room temperature W of 9.5 meV for the HRS state. Here we assumed the decay length α to be $0.2nm^{-1}$, a value commonly used disordered films [17-19], The R and W values obtained here are consistent with the constraints for VRH conduction that αR is equal to or less than unity and the hopping energy barrier is smaller than thermal energy. If otherwise the localization is strong enough (i.e. αR is larger than unity), nearest-neighbor hopping (NNH) process should be employed instead [16, 18].

The evolution of the conduction channels can be analyzed by studying how the distribution of V_Os evolve as the device is programmed from the HRS to an intermediate state then eventually to the LRS. To create the intermediate state, a 1 kΩ series resistor was attached in front of the device during programming, as shown in inset of Fig. 3.4(c). During programming the series resistor creates a voltage divider effect that slows down the filament growth as the RRAM resistance is reduced [7, 20-22]. As a result, an intermediate state between the HRS and the fully programmed LRS can be obtained. The intermediate state shows similar temperature dependence as the HRS, as shown in Fig. 3.4(d). Following the same treatment as the HRS, a density of localized (Vo) state of 2.0×10^{25} $eV^{-1}cm^{-3}$, average hopping distance of 0.24 nm and hopping energy of 0.83meV can be obtained. The significantly increased density of states for V_O and reduction of the hopping distance between V_O sites are consistent with the oxygen-vacancy mediated resistive switching model: as the device state changes from HRS to LRS, more oxygen

vacancies are accumulated in the conduction channel region and the distance between oxygen vacancies is significantly reduced with increased density of localized V_O states which in turn lead to a decrease in device resistance. These processes are schematically shown in Fig.

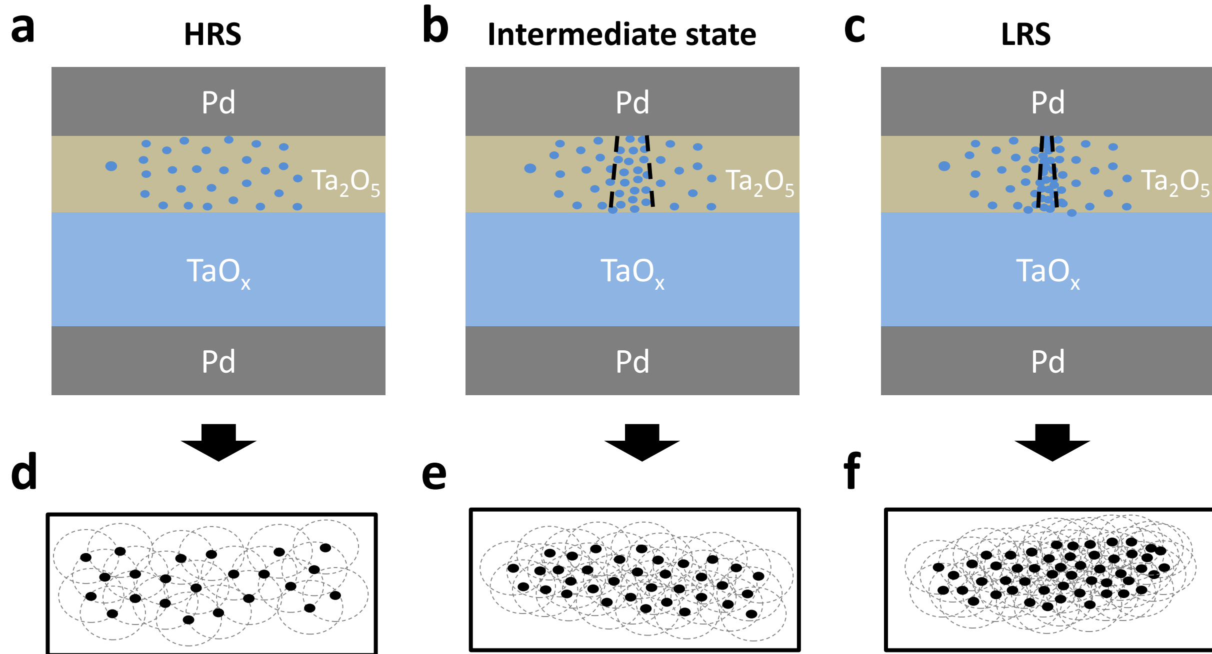


Figure 3.5. (a) Schematics showing the changes in V_O distribution for the HRS, the intermediate state and the LRS in the Ta_2O_5 switching layer, respectively. The dashed lines in (b) and (c) highlight the filament region with higher V_O concentration than the rest of the film. (d-f) Corresponding changes in the overlap of electron wavefunctions lead to different resistance values for the HRS, intermediate state and LRS. dots: localized states, Gray dashed circle: the localization radii.

3.5(a),(b).

3.5 Analysis based on noise and transport data

Based on noise and transport analyses, a unified picture of the resistive switching behavior can be obtained. Even in the HRS, the conduction through the device is not homogenous but rather dominated by "channels" with higher V_O concentration than the rest of

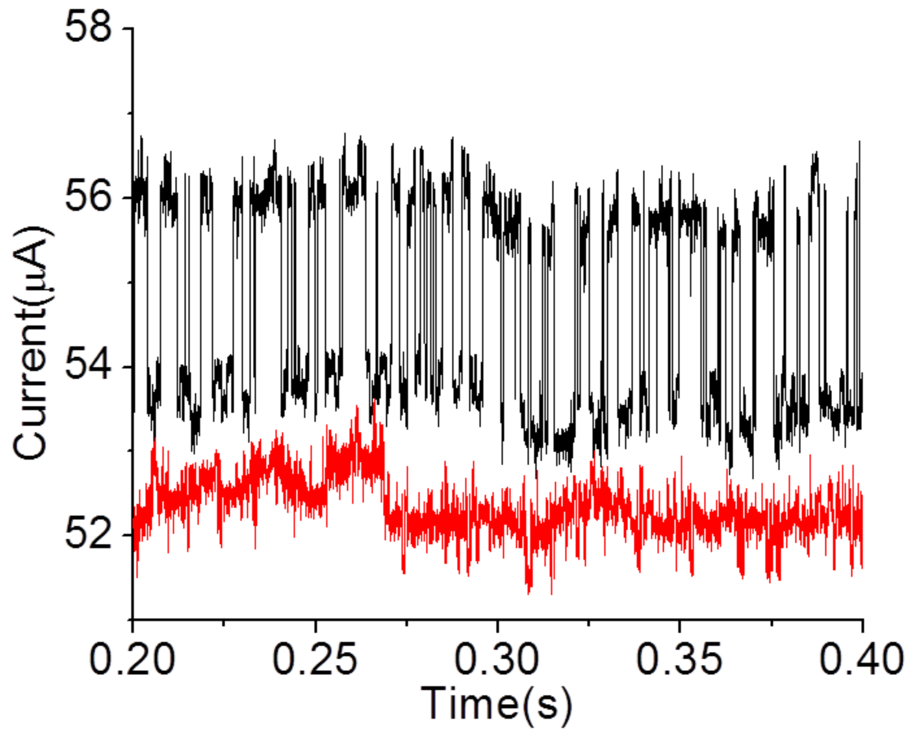


Figure 3.6. Current-time plots measured at 0.1 V on the same device after two different set and reset process.

the matrix, as shown in Fig. 3.5(a). Conduction through the channels is by electron hopping mediated by the V_O traps. While many V_O trap sites are distributed close to the average hopping distance of ~ 2 nm and form the conduction channel. Additionally, the broad V_O distribution in the film means that there will unavoidably be some V_O trap sites with distance much larger than the rest. These V_O traps are far enough from the channel to contribute to conduction current, and will rather act as a noise source as they occasionally trap and detrap electrons, as shown in Fig. 3.5(a). As a result, the RTN effect will be pronounced if only one or a few discrete V_Os are within the appropriate distance from the channel (too far the effect becomes very week while too close the traps become part of the channel itself [10]). These conditions are satisfied in the HRS when the V_O concentration is still not too high. This hypothesis also explains the observation that the RTN noise can vary significantly after each set and reset process even if the device is reset to

the same resistance value as shown in Fig. 3.6, since the distribution of V_Os will be different after each set and reset process.

We now turn to the switching process from the HRS to LRS. During this set process, more V_Os are accumulated in the conduction channel and the distance between the V_O sites becomes significantly reduced (as evidenced by the reduction of average V_O spacing from ~ 2 nm in the HRS to ~ 0.24 nm in the intermediate state) [23] and eventually extended electron states are formed when the electron wavefunctions overlap sufficiently, as shown in Fig. 3.5(f). This leads to the formation of metallic LRS states and the observed negative temperature coefficients. Significantly, our analysis also suggests the formation of LRS conduction channels is not uniform but rather a localized effect. Estimations of the effective channel size from the calculated conductivity using Eq. 5 and the measured conductance values resulted in a channel diameter of 42 nm for the HRS and only 3 nm for the intermediate state. Here we assumed a υ_{ph} value of 10^{12}/s [16] and channel length 5nm- which corresponds to the deposited Ta_2O_5 film thickness. What is interesting is that the channel (conduction filament) diameter for the intermediate state is much smaller than that of HRS, even though during resistance switching more V_Os are injected into the Ta_2O_5 layer. This result clearly suggests the conducting channel (filament) formation is a localized process, as non-uniformities in the V_O distribution in the film leads to a few local "hot spots" that attracts higher concentration of V_Os than the rest of the film. The enhanced local conductivity likely leads to higher local temperature which further speeds filament growth at these locations.

This filament formation picture is schematically illustrated in Figs. 3.5(a-c), which show the evolutions of oxygen vacancy distribution from the HRS through the intermediate state and

to the LRS; while Figs. 3.5(d-f) show how the conduction evolves from VRH to the formation of extended states. The lack of RTN in LRS can now be readily explained by the fact that there are simply too many V_O trapping sites whose noise effects overlap, making effects from individual trap sites (RTN) noise impossible to resolve, i.e. similar to the observation of the transition from RTN to regular 1/f noise in silicon inversion layers as the device size increases [10].

3.6 Conclusion

In summary, we found significant RTN exists in the HRS of the RRAM device. Systematic analysis of the RTN and electrical transport through the RRAM device verified the conduction channel formation are associated with the (re)distribution of V_Os. While in HRS the discrete V_Os outside the channel region can lead to significant RTN up to 25%, in LRS the higher V_O concentration causes the individual effects to overlap and the disappearance of RTN effects. Modeling of the transport data also leads to insight into the spacing of the V_O sites and the effective filament size. Specifically, as the device is switched from HRS to LRS, the conduction channel area is actually reduced with a much higher V_O density is obtained locally. These findings will provide valuable information on the application and design of oxide-based resistive switching devices for memory and logic applications.

3.7 References

[1] M.-J. Lee, C. B. Lee, D. Lee, S. R. Lee, M. Chang, J. H. Hur, Y.-B. Kim, C.-J. Kim, D. H. Seo, S. Seo, U.-I. Chung, I.-K. Yoo and K. Kim, Nat. Mater., 2011, 10, 625-630.

[2] J.-J. Yang, M.-X. Zhang, J. P. Strachan, F. Miao, M. D. Pickett, R. D. Kelley, G. Mederos-Ribeiro and R. S. Williams, Appl. Phys. Lett., 2010, 97, 232102.

[3] A. C. Torrezan, J. P. Strachan, G. Medeiros-Ribeiro, R. S. Williams, Nanotechnology, 2011, 22, 485203.

[4] B. Govoreanu, G. Kar, Y. Chen, V. Paraschiv, S. Kubicek, A. Fantini, I. Radu, L. Goux, S. Clima and R. Degraeve, et al., Electron Devices Meeting (IEDM), 2011 IEEE International, 2011, pp. 31–36.

[5] Y. Yang, P. Sheridan, W. Lu, Appl. Phys. Lett., 2012, 100, 203112.

[6] W. Lu, Z. Ji, L. Pfeiffer, K. W. West, A. J. Rimberg, Nature, 2003, 423, 422-425.

[7] S. H. Jo, K.-H. Kim, W. Lu, Nano Lett., 2009, 9(2), 870–874.

[8] R. Soni, P. Meuffels, A. Petraru, M. Weides, C. Kugeler, R. Waser, H. Kohlstedt, J. Appl. Phys., 2010, 107, 024517.

[9] D. Ielmini, F. Nardi, C. Cagli, Appl. Phys. Lett., 2010, 96, 053503.

[10] K. S. Ralls, W. J. Skocpol, L. D. Jackel, R. E. Howard, L. A. Fetter, R. W. Epworth, D. M. Tennant, Phys. Rev. Lett., 1984, 52, 228-231.

[11] H. D. Xiong, W. Wang, Q. Li, C. A. Richter, J. S. Suehle, W. K. Hong, T. Lee, D. M. Fleetwood, Appl. Phys. Lett., 2007, 91, 053107.

[12] Y. H. Tseng, W. C. Shen, C. J. Lin, J. Appl. Phys., 2012, 111, 73701-737015.

[13] J.-K. Lee, J.-W, Lee, J. Park, S.-W. Chung, J. S. Roh, S.-J. Hong, I.-W. Cho, H.-I. Kwon, J.-H. Lee, Appl. Phys. Lett., 2011, 98, 143502.

[14] J. J. Yang, D. B. Strukov and D. R. Stewart, Nat. Nanotechnol., 2012, 8(1), 13–24.

[15] S. Yu, X. Guan, H.-S. P. Wong, Appl. Phys. Lett., 2011, 99, 063507.

[16] N. Mott, E. Davis, Electronic Processes in Non-crystalline Materials; Clarendon press: Oxford, 1979.

[17] Z. H. Khani, M. M. Malik, M. Zulfequar, M. Husain, J. Phys. Condensed Matter, 1995, 4, 8979.

[18] I. Goldfarb, F. Miao, J. J. Yang, W. Yi, J. P. Strachan, M.-X. Zhang, M. D. Pickett, G. Medeiros-Ribeiro, R. S. Williams, Appl. Phys. A, 2012, 107, 1-11.

[19] M. Brodsky, R. Gambino, J. Non-Crystalline Solids, 1972, 810, 739-744.

[20] A. Fantini, D. Wouters, R. Degraeve, L. Goux, L. Pantisano, G. Kar, Y. Y. Chen, B. Govoreanu, J. Kittl, L. Altimime, et al., Memory Workshop (IMW), 2012 4th IEEE International. 2012, 1- 4.

[21] S. H. Jo, W. Lu, Nano Lett., 2008, 8(2), 392-397.

[22] K. H. Kim, S. H. Jo, S. Gaba, W. Lu, Appl. Phys. Lett., 2010, 96, 053106.

[23] F. R. Allen, C. J. Adkins, Philosophical Megazine, 1972, 26, 1027 - 1042.

Chapter 4.

Retention Failure Experiments and Modeling of Metal-Oxide Based RRAM

4.1 Introduction

In the previous two chapters, we discussed the basic picture of filament formation process through physics-based modeling, and performed systematic investigation of the resistance switching mechanism in a TaO_x based RRAM through detailed noise analysis. Both approaches show the resistance switching from high-resistance to low-resistance is accompanied by a semiconductor-to-metal transition mediated by the accumulation of oxygen-vacancies in the conduction path based on electrons transport. However, questions central to the device operation such as the switching and retention failure mechanism and important parameters such as the activation energy for oxygen vacancy (V_O) migration remain unsolved. By analyzing how the devices fail at elevated temperatures, we will not only confirm the switching mechanism of the devices as filamentary in nature but also be able to extract important device parameters such as the V_O activation energy from the failure time analysis.

4.2 Device Fabrication and Measurement Setup

The device studied here is based on the tantalum-oxide bilayer structure which consists of a Ta_2O_5 switching layer and an oxygen-deficient TaO_x base layer (Fig. 4.1(a)) [1-5]. The 40 nm

TaO_x base layer acts as a supply of oxygen vacancies for conductive filament formation and the 5 nm Ta_2O_5 layer is used as the switching layer where the conductive filament is formed and ruptured (Figure 4.1(a)) leading to resistance changes. The device fabrication starts with a Si/SiO_2 substrate with a 100 nm thermal SiO_2 layer. The bottom electrode (BE) consisting of 5 nm-thick NiCr and 40 nm-thick Pd was fabricated through photolithography and lift-off. The TaO_x layer was sputtered by direct current (DC) reactive using a Ta metal target with Ar/O_2 (32.3 SCCM/1 SCCM) gas mixture at 400 °C. Next, the Ta_2O_5 switching layer was deposited by radio frequency (RF) at room temperature using a Ta_2O_5 ceramic target. The base pressure of sputter chamber for both TaO_x and Ta_2O_5 layers was maintained under $\sim 10^{-6}$ Torr. To ensure the high quality of the interface between TaO_x layer and Ta_2O_5 layer, the films were deposited without breaking the vacuum. Finally, 40 nm of Pd and 20 nm of Au were patterned and deposited as the top electrode (TE) through photolithography and lift-off processes. The BE and TE were fabricated in a crossbar structure, as shown in the inset of Figure 4.1(b). Devices with size of 1 μm × 1 μm were fabricated and tested. During testing, the bias voltage was applied to the TE with the BE grounded. To perform the high-temperature retention analysis, a custom-built high temperature measurement setup was configured from a tube furnace (Carbolite, model CTF 12/75/700) with electrical feedthroughs to allow in-situ measurements at elevated temperature as shown in Figure 4.1(c). The devices were wirebonded to a chip carrier and connected to the electrical feedthroughs inside the furnace via wires wrapped by nonporous high-alumina ceramic wire tube as shown in Figure 4.1(d). The data are collected by a custom data acquisition system and a DL 1211 current preamplifier from DL industries.

4.3 Experimental Data

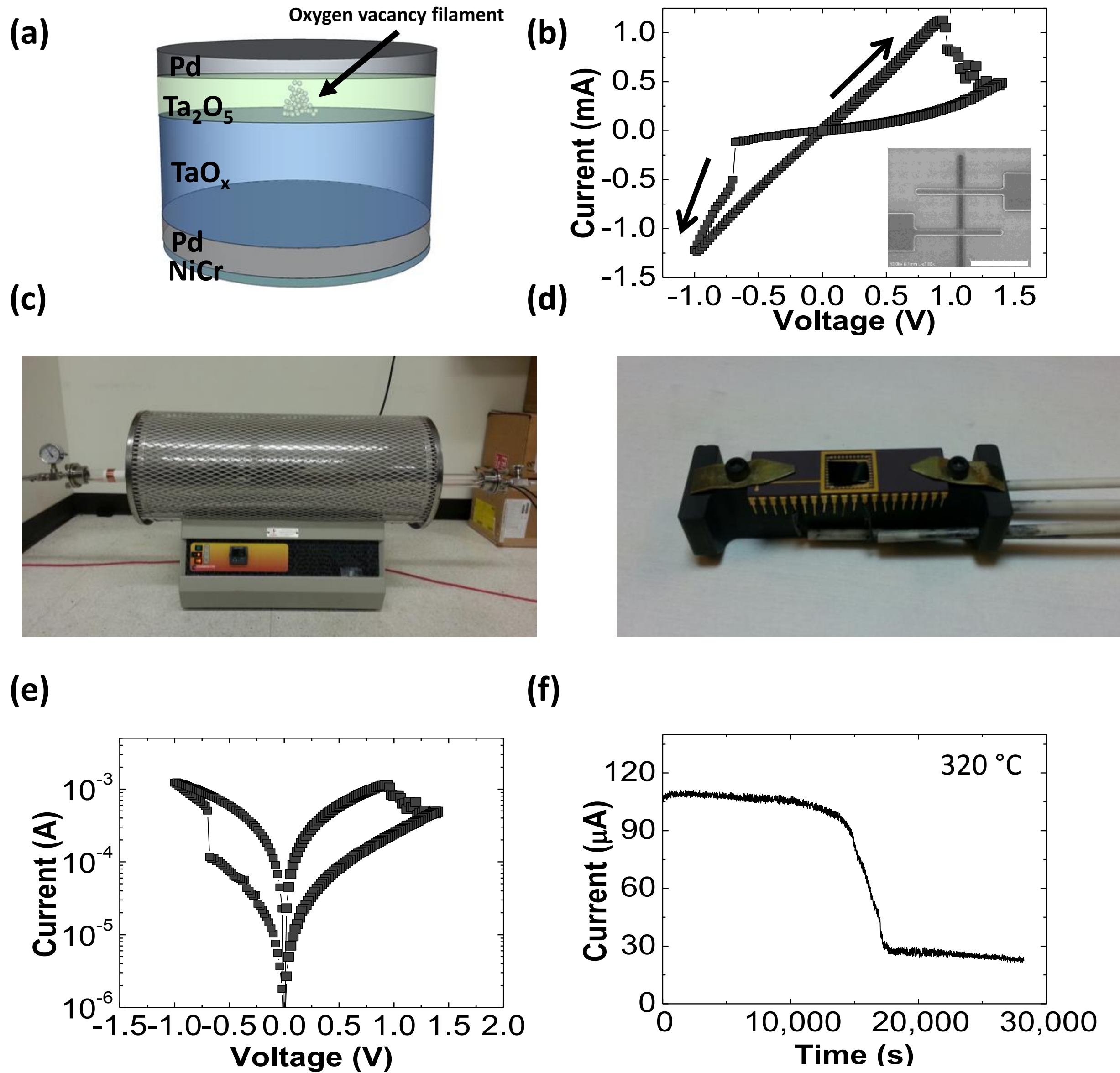


Figure 4.1. (a) Schematic of the Pd/Ta_2O_5/TaO_x/Pd bilayer RRAM device. (b) DC I-V characteristics of the device showing the bipolar switching behavior. Inset: SEM image of the device. Scale bar is 20 μm. (c) A custom-built high temperature measurement setup using a tube furnace. The left part of the tube is connected to a vacuum pump and the right part of the tube is connected to electrical feedthroughs. (d)The wirebonded devices on a chip carrier in the furnace connected to electrical feedthroughs. (e) I-V characteristics of the device in log scale. (f) Retention measurement results at 320 °C. A read pulse (0.1 V/10 ms) was applied every 6s during the test.

The device shows typical bipolar resistive switching behavior (Fig. 4.1(b), (e)) and can be SET to a low-resistance state (LRS) with a negative voltage and RESET to a high-resistance state (HRS) with a positive voltage. As discussed in previous chapters, the resistive switching is believed to be caused by oxygen vacancy (V_O) redistribution and the formation of V_O-rich conductive filaments in the Ta_2O_5 layer [6-9]. Below we show this model can be used to quantitatively explain the retention failure of the RRAM devices, and can be used to extract important microscopic physical parameters such as the oxygen vacancy migration activation energy through simple temperature dependent measurements.

A typical retention failure is shown in Fig. 4.1(f). Here the device conductance in LRS was periodically monitored at 320 ℃ (593 K) in every 6s with a low read voltage pulse (0.1 V/10 ms) to avoid disturbance of the device state. As shown in Fig. 4.1(f), an initial slow conductance drift followed by an abrupt conductance drop were normally measured, with the abrupt drop in conductance corresponding to the rupture of the conductive filament.

Temperature dependent studies were carried out to reveal the nature of the filament failure, as shown in Fig. 4.2(a). As expected, the device failed faster at higher temperatures. Additionally, two different regimes of conductance change, the gradual drift and the fast drop, were observed at all temperatures. By measuring the time at the intersection between the gradual and the sudden change, the retention failure time can be determined. The retention failure time as a function of temperature was recorded and analyzed as shown in Fig. 4.2(b), which shows an apparent thermal activation effect with an activation energy of 1.6 eV. Significantly, extracting the retention time based on these high temperature data yields retention of 1.6×10^{10} years at

room temperature and 9.7×10^5 years at 80 ℃, verifying the excellent retention property of the tantalum oxide based RRAM devices.

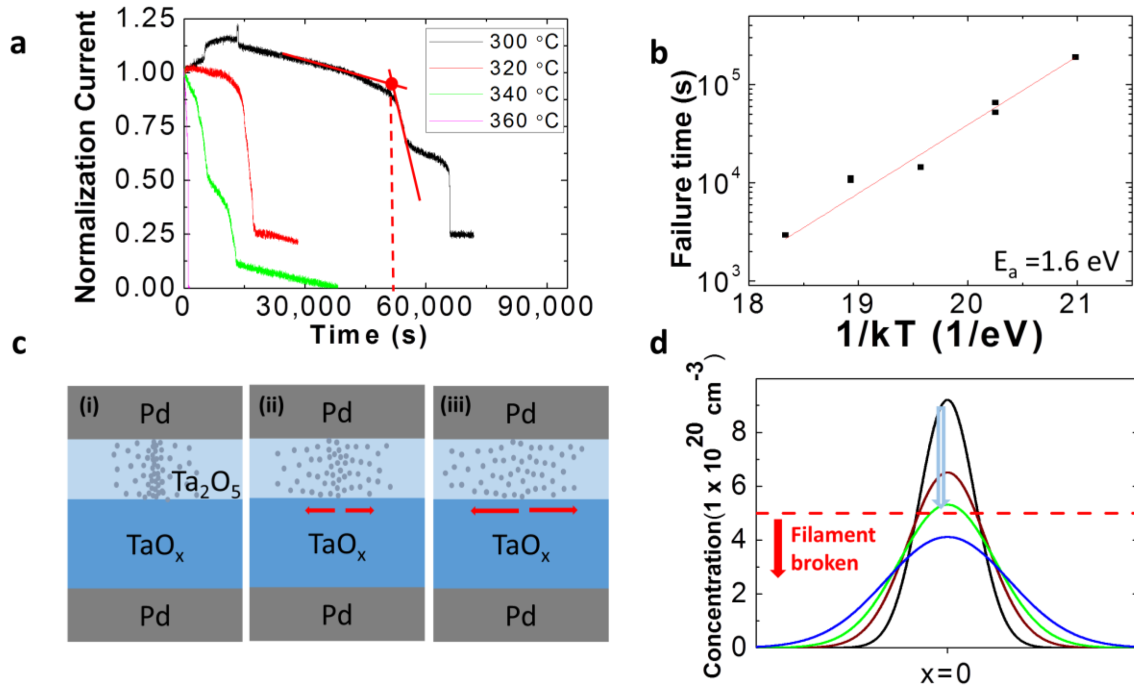


Figure 4.2. (a) Temperature dependent retention measurements at 300 ℃, 320 ℃ , 340 ℃ and 360 ℃. (b) Temperature dependence of the characteristic retention failure time (squares) and fitting (line) following the Arrhenius equation. (c) Schematics showing the changes in V_O distribution from the LRS (i), after VO out diffusion (ii), and eventual rupture of the filament (iii). (d) Oxygen vacancy concentration profile predicted from the simple analytical model as a function of time. Dashed red line indicates the critical oxygen vacancies density (defined as $5\times 10^{20}/cm^3$).

The next question to be addressed is how the measured activation energy is related to the microscopic physical parameters of the device. The retention loss can be understood from the V_O-based filament model as shown in Figure 4.2(c). A conducting filament is formed when enough V_Os are accumulated in the region, leading to LRS (i). The V_Os inside the filament, however, can be diffused away through spontaneous diffusion, which is a thermally activated process (ii). As a result, the V_O concentration inside the filament is gradually reduced, corresponding to the initial drift of device conductance. Finally, over time the V_O concentration

inside the filament is reduced below a critical value, i.e., when the electron wavefunctions associated with the V_Os no longer overlap and an extended state is no longer formed [10,11], the filament is effectively ruptured corresponding to an abrupt drop in device conductance.

To verify this model and extract important device parameters, we performed analytical calculations of the V_O concentration inside the filament based on the V_O diffusion model. The V_O distribution was assumed to be cylindrically symmetric and follow a Gaussian distribution along the radial direction [12]. The V_O concentration at position x away from the center of the filament at given time t and temperature T can be written as,

$$\mathrm{N(x,t)} = \frac{N_O}{\sqrt{\pi D t}} \exp\left(-\frac{x^2}{4Dt}\right) \tag{1}$$

where N_O is total number of V_Os scaled over the device area and D is the diffusion coefficient. Here we neglected the background V_O concentration in the Ta_2O_5 layer. The retention time, t_c, corresponding to the time when the oxygen vacancy peak concentration (at x=0) becomes smaller than the critical V_O density, N^*, can then be calculated as

$$\mathrm{N(x=0, t_C)} = \frac{N_O}{\sqrt{\pi D t_c}} = N^*. \tag{2}$$

Here, the diffusion coefficient D can be written as

$$\mathrm{D} = D_O \exp\left(-\frac{E_a}{kT}\right) . \tag{3}$$

where E_a is the activation energy for V_O migration. By plugging Eq. (3) in to Eq. (2), the relation between the retention time, t_c, and temperature, T, can be obtained

$$t_c = \frac{1}{\pi D_0} \left(\frac{N_0}{N^*}\right)^2 \exp\left(\frac{E_a}{kT}\right) . \tag{4}$$

From (4), one can see that the retention time t_c indeed shows a thermal activation behavior. More importantly, we show that the extracted activation energy E_a from the simple

retention measurement in fact corresponds to the microscopic activation energy for V_O migration, which is an important physical parameter needed for the design and optimization of oxide based RRAM devices.

4.4 **Multiphysics Simulation**

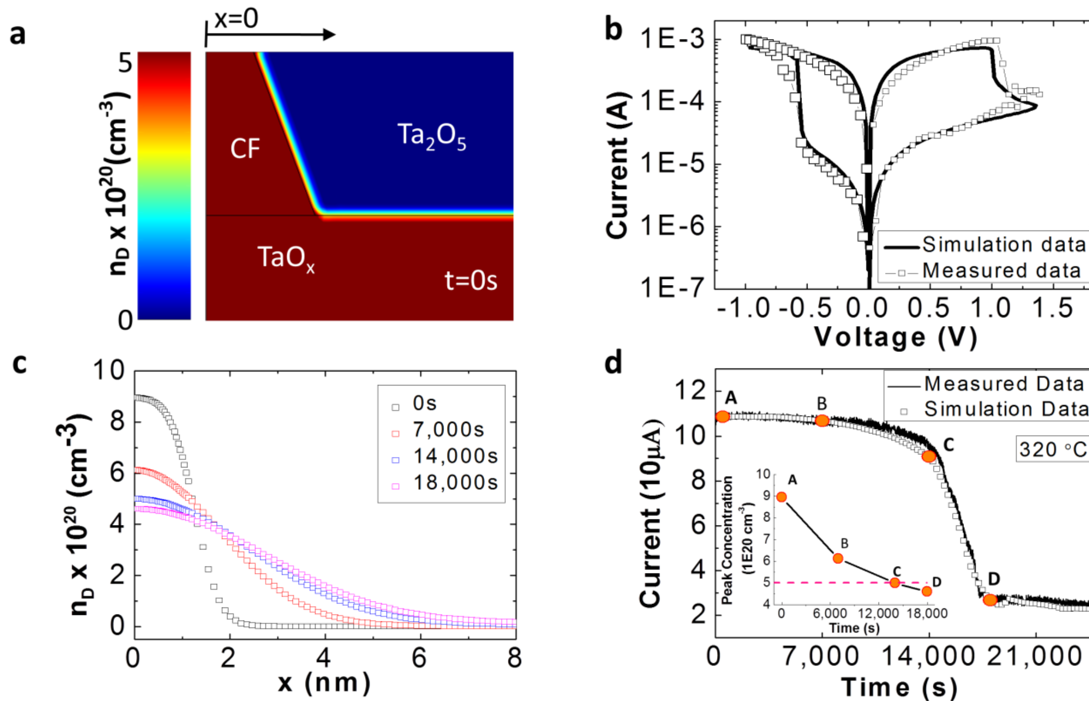


Figure 4.3. (a) 2-D map of oxygen vacancy concentration obtained through numerical simulations, for in the initial state (LRS). The x=0 position is the center of the conductive filament. (b) Measured and calculated DC I- V characteristics of the device at 320 °C showing the model can capture the essential dynamic V_O migration processes. (c) Oxygen vacancy concentration profile calculated from the numerical simulation as a function of time. (d) Measured and simulation results showing the device retention behavior at 320 °C inset: Peak V_O concentration at different time instants (A- D). Dashed red line indicates the critical oxygen vacancies density (defined as 5× $10^{20}/cm^3$).

Results from the simple analytical model were confirmed through detailed numerical simulations by self-consistently solving the drift and diffusion continuity equation for oxygen vacancy transport, current continuity equation for electron transport and joule heating effects [6-9]. Figure 4.3(a) shows a 2-D map in the axial and radial plane of the V_O concentration at the

LRS state during the simulation. The filament size at the top part of the switching layer is 2~ 3 nm as calculated in Ref. [10]. Figure 4.3(b) shows the measured and calculated DC I-V characteristics during set and reset processes at 300 ℃. The reset and set transition occurs at 1 V and -0.5 V, respectively. The simulation quantitatively captures the resistive switching dynamics with a fixed set of material specific parameters [10], and confirmed the parameters are appropriately chosen. Figure 4.3(c) shows the simulation result of the V_O concentration at the topmost region of the conductive filament as a function of time during retention test at 320 ℃. The detailed simulation results are consistent with the simple analytical model results shown in Fig. 4.2(d) and support the V_O diffusion model in device retention failure analysis. Additionally, the conductance of the device can be directly obtained from the simulation [10], and Fig. 4.3(d) plots the measured and calculated conductance during the retention measurement. Again the retention behavior can be quantitatively predicted through simulation and verifying the accuracy of the filamentary model. The inset of Figure 3d shows the peak V_O concentration at 0 s, 7,000 s, 14,000 s and 18,000 s corresponding to the yellow points in Figure 4.3(d) and the dots in Figure 4.3(c). The dashed line indicates the critical V_O concentration which is chosen as 5×10^{20} cm^{-3} from literature [6, 7] and corresponds to the point C in the retention curve in Fig. 4.3(d). As expected the conductance of the device decreases abruptly beyond this point as the filament is considered ruptured beyond this point.

4.5 **Monotonic Current Increase Behavior**

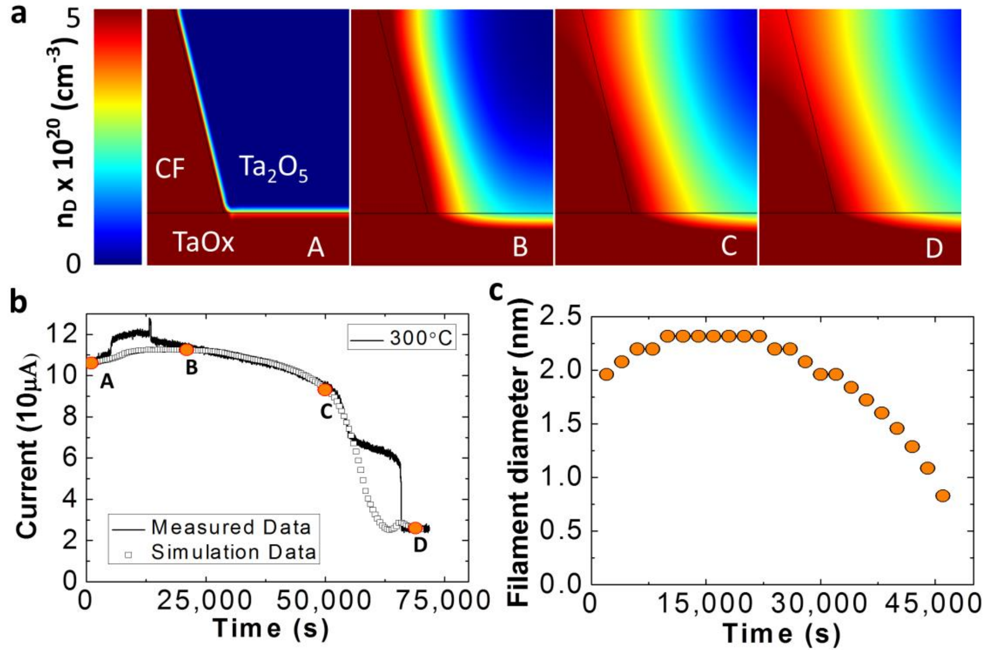


Figure 4.4. (a) 2-D map of oxygen vacancy concentration showing the evolution of the filament at different time scales (corresponding to points A-D in Figure 4b) (b) Measured and calculated device conductance as a function of time at 300 °C. (c) Effective filament diameter as a function of time. The filament was defined as the region with oxygen vacancy concentration higher than 5× 10^{20} cm^{-3}.

Interestingly, it can be found that at points A and B, the device conductance does not change much even though the peak oxygen vacancy concentration decreases from 9× 10^{20} cm^{-3} to 6.1× 10^{20} cm^{-3}. This can be explained by the fact that the effective filament area is increased accompanying the decrease in peak V_O concentration due to out diffusion (e.g. Figs. 4.2(d), 4.3(c)). The increase in effective filament size counters the effect of peak V_O concentration decrease. This effect is more pronounced by examining the retention test obtained at 320 °C, replotted in Fig. 4.4(b). Surprisingly, an initial increase, instead of decrease of conductance, is observed from point A to B. The conductance eventually decreases and abruptly drops at point C, corresponding to the rupture of the filament. This non-monotonic behavior is somewhat

counterintuitive but can be fully explained by the model discussed earlier. For example, Figure 4.4(a) plots the 2D map of the V_O concentration in the filament at different time instants, and an effective increase of the filament size is clearly observed (from A to B), accompanying the decrease of the peak V_O concentration due to diffusion. The non-monotonic change in conductance during retention measurement was again quantitatively predicted by the model, as shown in Fig. 4.4(b). Figure 4.4(c) shows the calculated effective filament diameter (defined as the region with V_O concentration > 5×10^{20} cm^{-3}). It clearly shows that the effective diameter of the filament increases initially and eventually starts to decrease and finally break over time.

4.6 Multi-filament Effects

Finally, we show that the experiments provide evidence for multiple filaments in some cases. The solid line in Figure 4.5 plots the retention test results at 340 ℃. After the first abrupt drop in conductance at 3,700 s, the conductance did not drop to the background level, but rather reached a second plateau at 6,000s, and eventually reached to the background level after a second abrupt drop at 10,000 s. The experimental results can be quantitatively explained by the simulation data (dots), by considering the evolution of two filaments, as shown in Figure 4.5. The initial filament diameters for the two filaments were chosen as 2.3 nm to 2.7 nm, respectively, with all other parameters remaining the same. The excellent agreements between simulation results and experimental data not only prove again the quality of the model, but also help shed light into future design and optimization of this important class of devices.

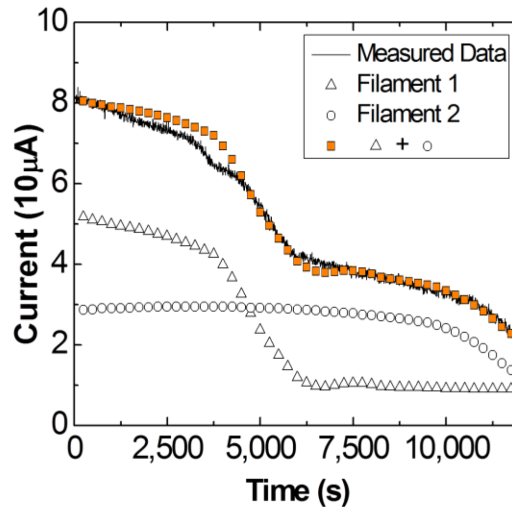


Figure 4.5. Measured (black line) and calculated (squares) conductance as a function of time at 340 °C, showing the possible existence of multiple filaments. Evolutions of the two filaments (triangles and circles) were obtained through simulation and the overall conductance (squares) is the sum of the two filaments.

4.7 **Conclusion**

In summary, we analyzed retention behaviors of oxide-based RRAM at elevated temperatures and matched the experimental results with an oxygen vacancy diffusion model. Our analysis shows that the activation energy for oxygen vacancy migration can be directly calculated from the failure time versus temperature relationship. A non-monotonic conductance drift was also observed and can be explained within the oxygen vacancy out diffusion framework. Evidence for multiple filaments was also examined and supported by simulation. These findings support the filamentary model of RRAM devices and shed valuable insight in the design and optimization of oxide-based resistive switching devices for memory and logic applications.

4.8 References

[1] M.-J. Lee, C.B. Lee, D. Lee, S.R. Lee, M. Chang, J.H. Hur, Y.-B. Kim, C.-J. Kim, D.H. Seo, S. Seo, U.-I. Chung, I.-K. Yoo, and K. Kim, Nat. Mater. 10, 625 (2011).

[2] J.J. Yang, M.-X. Zhang, J.P. Strachan, F. Miao, M.D. Pickett, R.D. Kelley, G. Medeiros-Ribeiro, and R.S. Williams, Appl. Phys. Lett. 97, 232102 (2010).

[3] A.C. Torrezan, J.P. Strachan, G. Medeiros-Ribeiro, and R.S. Williams, Nanotechnology 22, 485203 (2011).

[4] J.P. Strachan, A.C. Torrezan, G. Medeiros-Ribeiro, and R.S. Williams, Nanotechnology 22, 505402 (2011).

[5] Y. Yang, P. Sheridan, and W. Lu, Appl. Phys. Lett. 100, 203112 (2012).

[6] S. Kim, S. Choi, and W. Lu, ACS Nano 8, 2369 (2014).

[7] S. Kim, S.-J. Kim, K.M. Kim, S.R. Lee, M. Chang, E. Cho, Y.-B. Kim, C.J. Kim, U.-I. Chung, and I.-K. Yoo, Sci. Rep. 3, 1680 (2013).

[8] F. Nardi, S. Balatti, S. Larentis, and D. Ielmini, Tech. Dig. – Int. Electron Devices Meet. 2011, 709.

[9] S. Larentis, F. Nardi, S. Balatti, D.C. Gilmer, D. Ielmini, IEEE Trans. Electron Devices 59, 2468 (2012).

[10] S. Choi, Y. Yang, and W. Lu, Nanoscale 6, 400 (2014).

[11] I. Goldfarb, F. Miao, J.J. Yang, W. Yi, J.P. Strachan, M.-X. Zhang, M.D. Pickett, G. Medeiros-Ribeiro, and R.S. Williams, Appl. Phys. A 107, 1 (2012).

[12] Z. Wei, T. Takagi, Y. Kanzawa, Y. Katoh, T. Ninomiya, K. Kawai, S. Muraoka, S. Mitani, K. Katayama, S. Fujii, R. Miyanaga, Y. Kawashima, T. Mikawa, K. Shimakawa, and K. Aono, T Aono, Tech. Dig. – Int. Electron Devices Meet. 2011, 721.

Chapter 5.

Tuning Resistive Switching Characteristics of Tantalum-Oxide RRAM Devices through Si Doping

5.1 **Introduction**

As we discussed in the previous chapters, the RS behavior is believed to be caused by the transport of oxygen ions (O^{2-}) and consequent oxygen vacancy (V_O) redistribution in the oxide layer, where high V_O concentration regions (*e.g.*, conducting filaments (CFs)) provide high conductance channels for electrical transport [1]. The device can be set (from a high resistance state (HRS) to a low resistance state (LRS)) or reset (from LRS to HRS) between the different resistance states according to the formation/rupture of the CFs. Although a single oxide layer can attain this RS behavior in RRAM devices [2-4], bi- [5,6], triple- [7], or even quadruple- [8] layered oxides have been explored in recent years to improve the switching characteristics. The additional oxide layers act as O^{2-} or V_O reservoirs, and can improve the device reliability (*e.g.*, cycling endurance or switching uniformity) by confining RS at selected layers [5,6]. In other cases, the additional oxide layer serves as a tunneling barrier, which induces current nonlinearity by suppressing the leakage current at the low-voltage regime [7]. However, although these extra layers improve the RS controllability, the device-to-device variation and fabrication complexity increases as the number of oxide layers increases, potentially affecting high-density device integration. Therefore, providing a fundamental atomic-level design that can directly control the

dynamic transport of ions within the switching layer not only allows tuning of the RS behavior but also significantly expands the parameter space for material and device optimization, which will be critical for continued development of RRAM devices.

In this chapter, we show that the RS dynamics in a tantalum-oxide-based bilayer RRAM can be modulated through doping of Si atoms in the Ta_2O_{5-x} switching layer. The additional dopant modifies the atomic structure and creates preferred V_O transport channels. Even a small amount of dopants can significantly affect the V_O drift process and change the ion hopping distance and drift velocity, thus allowing control of the RS process at the atomic level. The roles of the dopants were revealed through *ab initio* calculations and confirmed experimentally by extracting the effective V_O hopping distance through a series of measurements. Finally, we show the Si-doped tantalum-oxide bilayer RRAM devices can emulate different synaptic plasticity with excellent cycling endurance, and is suitable for future neuromorphic computing applications.

5.2 **Device Fabrication**

The RRAM devices used in this work with a size of 1 μm × 1 μm were fabricated in a crossbar structure on SiO_2 (100 nm)/Si substrates with electrodes patterned using traditional photolithography (GCA AS200 AutoStep). First, a 35-nm bottom Pd electrode was deposited by photolithography, e-beam evaporation, and lift-off processes. Next, a 30-nm TaO_x base layer was deposited by DC reactive sputtering of a Ta metal target in an Ar/O_2 gas mixture at 400 ℃. The total pressure of Ar/O_2 was ~5 mTorr, and the oxygen partial pressure in the Ar/O_2 mixture was 3%. A 5-nm Ta_2O_5 switching layer was then deposited by RF sputtering using a Ta_2O_5 ceramic target, and p-doped Si was co-deposited with Ta_2O_5 by DC sputtering (DC power was 40, 70 W,

and 140 W to achieve effectively 2.7%, 4.2% and 9.3% Si doping, respectively) in Ar with a pressure of ~5 mTorr. A 30-nm top Pd electrode was then deposited by photolithography, e-beam evaporation, and lift-off processes. Finally, a reactive ion etching process using SF_6/Ar was performed to expose the bottom contacts. For the forming step, a resistor (5 kΩ) was serially connected to the device to prevent permanent breakdown.

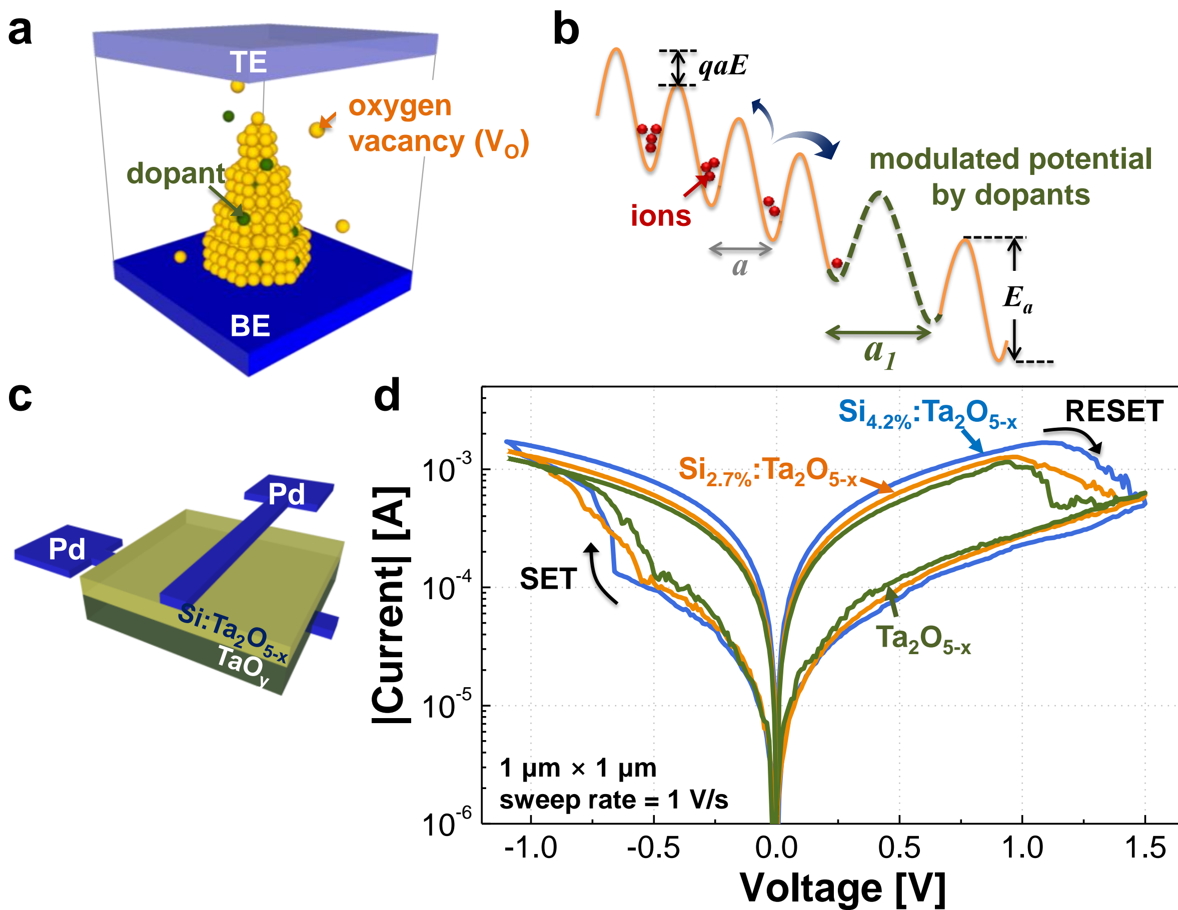


Figure 5.1. (a) Conceptual schematic of the oxide RRAM during RS. The agglomerated V_Os enhance the local electrical conductivity and form the CF. (b) Schematic of the potential energy landscape for ion hopping under electric field *E*. (c) Schematic plot of the Pd/Si:Ta_2O_5/TaO_x/Pd bilayer RRAM device. (d) DC *I*–*V* characteristics of undoped Ta_2O_5, $Si_{2.7\%}$:Ta_2O_5, and $Si_{4.2\%}$:Ta_2O_5 devices. The measured device size is 1 μm × 1 μm, and the voltage sweep speed is 1 V/s.

5.3 **Resistive Switching Behavior**

Figure 5.1(a) shows a conceptual schematic to explain the RS behavior in an oxide RRAM. The CF corresponds to the region with agglomerated V_O, which control the local electrical and thermal conductance properties [1,9]. The set and reset processes are described by the ionic transport and consequent V_O migration induced by the local electric field and temperature due to Joule heating. The nonlinear ionic transport under high electric field can be explained by the simple rigid-point-ion model shown in Figure 5.1(b) [10-12]. Oxygen ions (equivalently V_Os) hop among the energy potential wells (assumed to have hopping distance a and energy barrier E_a), where the applied electric field E lowers the energy barriers by a factor of qaE. The average V_O drift velocity is given as

$$v = a \cdot f \cdot \exp(-E_a/kT) \cdot \sinh(qaE/2kT) \quad (1)$$

where f, k, T, and q are the frequency of escape attempts, Boltzmann constant, temperature, and electron charge, respectively [1,12]. From this equation, one can see that the V_O drift is a strongly non-linear function of the applied electric field E through the sinh function, where the effect can also be strongly affected by the hopping distance a, which can in turn be affected by doping, as schematically shown in Figure 5.1(b). In other words, the RS behavior controlled by the ion hopping process can be fundamentally optimized at the atomic level through doping. To demonstrate this concept, a tantalum-oxide-based bilayer RRAM with Si dopants was fabricated, which consisted of a resistive Si-doped Ta_2O_5 film as the RS layer and a conductive TaO_x film as the V_O reservoir [5,7]. These two layers were sandwiched by top and bottom Pd electrodes (TE and BE), as shown in Figure 5.1(c). The Si dopants were introduced by a co-sputtering process

during the Ta_2O_{5-x} film deposition where the atomic percentage of Si was controlled by the sputtering power.

Figure 5.1(d) shows DC *I–V* characteristics in three different devices based on RS layers of undoped Ta_2O_5, $Si_{2.7\%}$:Ta_2O_5, and $Si_{4.2\%}$:Ta_2O_5, respectively, during the set and reset processes. Noticeable differences can be found when Si dopant is added: 1) the current level at the LRS increases, 2) the current level at the HRS decreases, and 3) the set voltage (defined as the voltage when the current begins to increase abruptly) increases. These different switching behaviors indicate that the formation/rupture of the CFs is modulated by the Si dopant. These effects are observed more clearly in pulse switching tests.

5.4 **Switching Dynamics Analysis**

Figure 5.2(a) shows the evolution of the device conductance as a function of the reset-pulse amplitude, by keeping the set-pulse amplitude constant while increasing the reset-pulse amplitude from 0.6 to 1.7 V during repeated set and reset processes. A more conductive LRS and a more resistive HRS are clearly obtained in the Si-doped devices. In a generally accepted theory [13], the difference in HRS and LRS can be explained by the evolution of the CF shape during switching: a thicker CF radius (r) inside the switching layer leads to higher current at the LRS

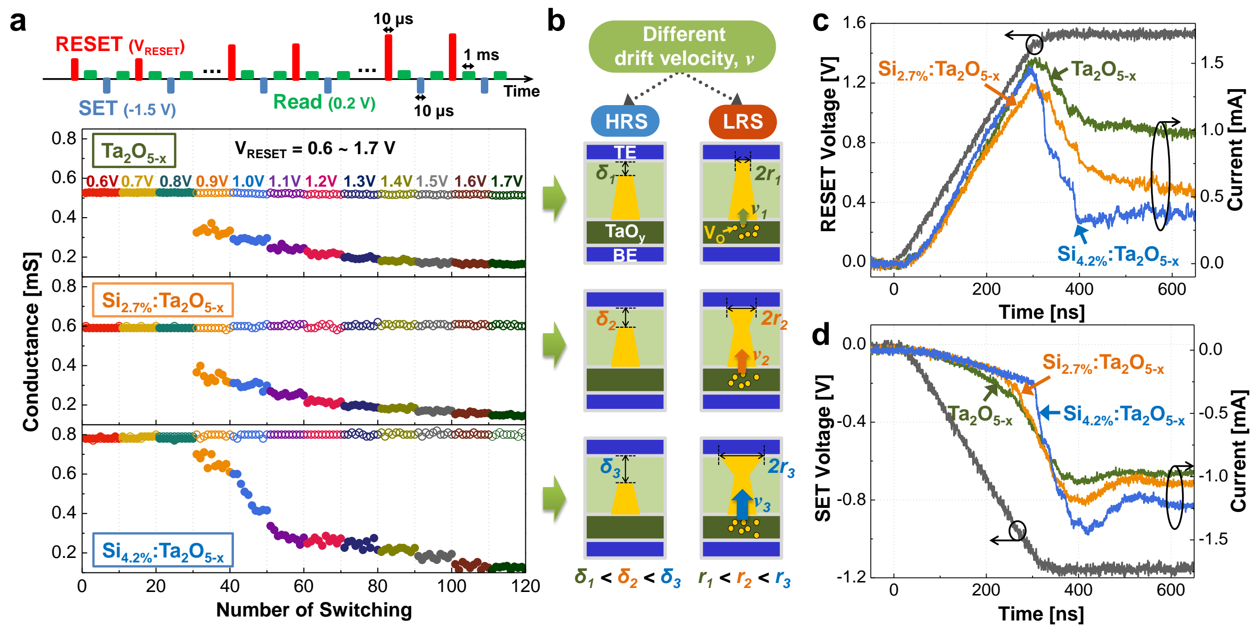


Figure 5.2. (a) Switching dynamics characterized by pulse measurements with increasing reset pulse amplitudes. The set-pulse amplitude is fixed at -1.5 V. Before the pulse measurements, the devices are set to LRS with a DC voltage sweep. (b) Schematic of the HRS and LRS for different V_O drift velocities. δ and r increase as v is increased through Si doping. (c) Time-dependent switching transient during the reset and (d) set processes.

(*i.e.*, more conductive LRS), and a wider depleted gap length (δ) between the TE and CF tip leads to lower current at the HRS (*i.e.*, more resistive HRS), as shown in Figure 5.2(b). These

results thus suggest that Si doping allows the V_Os to drift faster under the electric field during set/reset that leads to lower LRS resistance and higher HRS resistance. The V_O drift will also be accelerated by Joule heating effects [1,9,10]. For example, during the set process, a faster V_O drift (field-driven) leads to a fast decrease of device resistance, which results in enhanced Joule heating accelerates the V_O drift further [1,9,10]. During reset the lower LRS may also cause higher Joule heating and lead to a faster V_O drift [1,9,10] (thermally-accelerated) and consequently, a wider depleted gap. To accurately observe the time-dependent drift process, the transient response of the RS was measured, as shown in Figures 5.2(c) and 5.2(d). We observed the Si-doped devices show faster resistance transitions during set and reset processes, which again indicates a faster V_O drift process. Therefore, the different RS behaviors *via* Si doping can be understood by the faster V_O drift process: during set, the V_O can be easily supplied from the TaO_y layer under the same applied electric field leading to larger CF formation and a more conductive LRS; while during reset, the V_Os can be more easily depleted leading to a wider gap δ and a more resistive HRS, as shown in Figures 5.3(b).

5.5 Evaluation of the Effective Hopping Distance

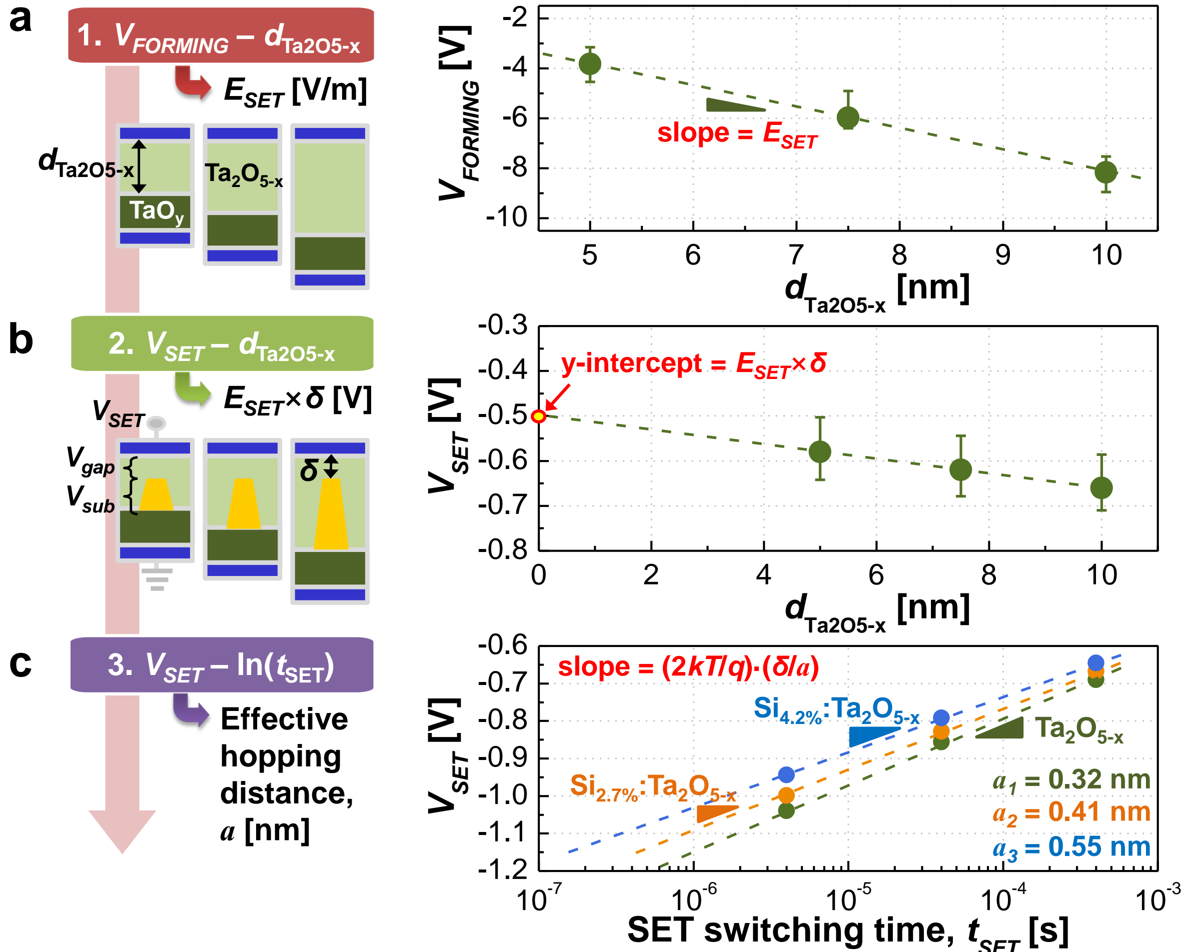


Figure 5.3. The three-step measurement procedure to evaluate the effective hopping distance. (a) $V_{FORMING}$–d_{oxide} plot to evaluate E_{SET}; the slope represents E_{SET}. (b) V_{SET}–d_{oxide} plot to estimate δ; the y-intercept represents $E_{SET}\cdot\delta$. (c) V_{SET}–ln(t_{SET}) plot to extract a; the slope represents $(2kT/q)\cdot(\delta/a)$.

As shown in Eq. (1), the difference in V_O drift velocity can be explained by the modulation of the hopping distance *a*. To estimate this effective hopping distance in the switching layer, we developed a method on the basis of a series of electrical measurements. Figure 5.3 shows a three-step measurement procedure to evaluate the effective hopping distance. First, Figure 5.3(a) shows the measured forming voltage ($V_{FORMING}$) according to different switching layer thicknesses (d_{oxide}) to evaluate the critical electric field (E_{SET}) that initiates V_O

migration. We assume that the entire forming voltage is applied on the switching layer, and the voltage drops in the much more conductive electrodes and the TaO_x layer are neglected. Then, E_{SET} can be estimated from the slope of the $V_{FORMING}$–d_{oxide} plot on the basis of the relationship $V_{FORMING} = E_{SET} \times d_{oxide}$. Second, V_{SET} as a function of different d_{oxide} values are measured to evaluate δ of the CF. When the set process is initiated by V_{SET}, the applied V_{SET} is divided into two parts, as shown in Figure 5.3(b), *i.e.*, $V_{SET} = V_{gap} + V_{sub}$. V_{gap} is applied across the gap δ and can be expressed as $V_{gap} = E_{SET} \times \delta$. We assumed that δ is mainly determined by the reset conditions and will be constant at a given reset condition regardless of d_{oxide}. On the other hand, because the length of the remnant CF (hence its resistance) is proportionally increased with increase of d_{oxide}, V_{sub} is to first order proportional to d_{oxide}. Thus, V_{gap} ($= E_{SET} \cdot \delta$) can be extracted from the y-intercept value of the V_{SET}–d_{oxide} curve; consequently, δ is estimated using the evaluated E_{SET} in the first measurement. The experimentally extracted δ is consistent with the estimated values by previous studies [9,14], and δ increases from 0.58 to 0.71 nm with the addition of the Si dopant consistent with the observed more resistive HRS behavior in the Si-doped devices shown in Figure 5.2(a). Finally, the effective hopping distance can be estimated from the slope of the V_{SET}–$\ln(t_{SET})$ curve, where t_{SET} is defined when the normalized conductance change ratio reaches two, and the slope is expressed as $(2kT/q)\cdot(\delta/a)$ which will be discussed in the next section. From these three-step measurements, the effective hopping distance can be quantitatively extracted, where a increases from 0.32 to 0.55 nm as the Si dopant is added, which contributes to a faster drift process, as predicted in Eq. (1) and explains the experimentally observed higher HRS and lower LRS in Si-doped devices shown in Figures 5.2(c) and 5.2(d).

5.6 Details of the model

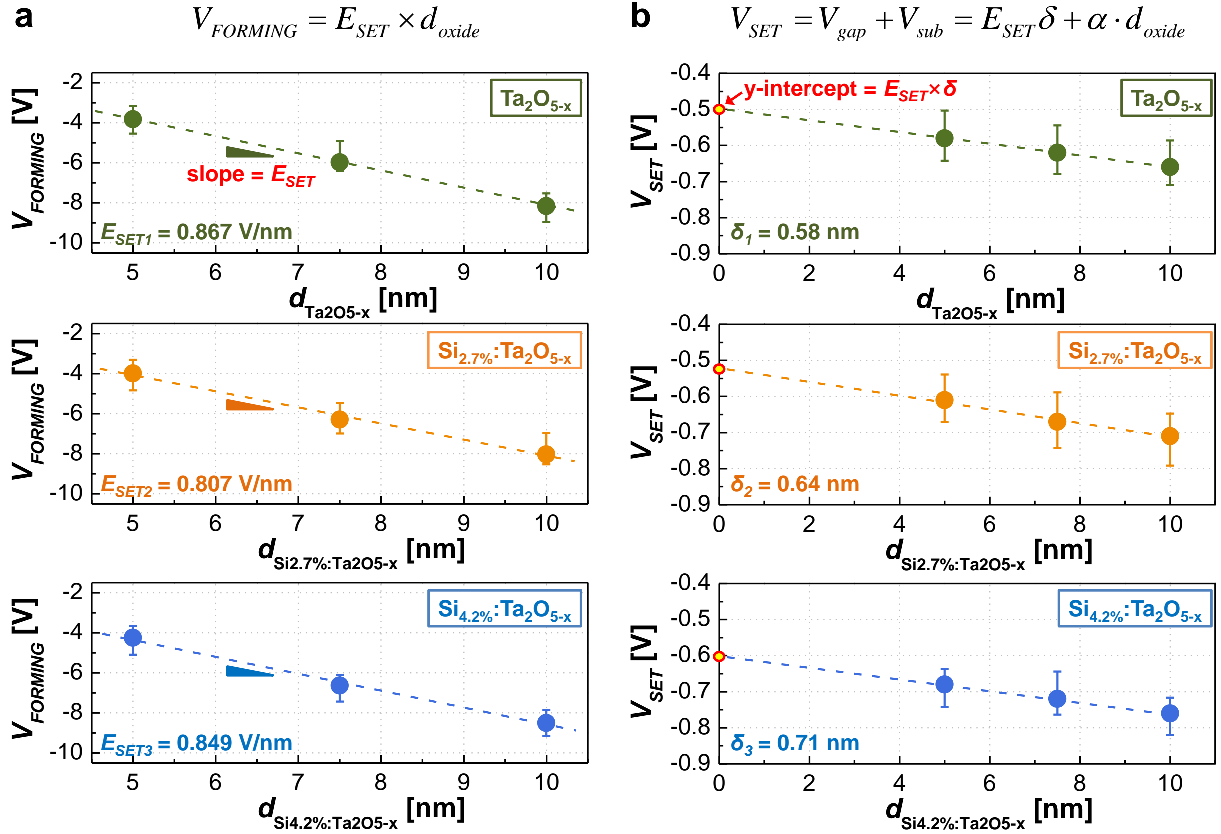


Figure 5.4. Measurement results of (a) $V_{FORMING}$–d_{oxide} to extract E_{SET} and (b) V_{SET}–d_{oxide} to extract δ in the three different samples.

Figure 5.4(a) shows the extracted E_{SET} data of the three different samples according to the $V_{FORMING}$–d_{oxide} relationship. We note that the extracted E_{SET} values are almost equal among the three different samples. In addition, the depleted gap lengths (δ) are estimated from the V_{SET}–d_{oxide} plot, as shown in Figure 5.4(b), where δ increases from 0.58 to 0.71 nm as the Si dopant is added. These results are consistent with the more resistive HRS in the Si-doped layer, as shown in Figure 2a.

Effective hopping distance (a) can be extracted from the slope of the V_{SET}–ln(t_{SET}) curve. The slope can be expressed as

$$slope = \frac{dV_{SET}}{d\ln t_{SET}} = \frac{dV_{SET}}{d\delta} \cdot \frac{d\delta}{d\ln t_{SET}} \quad (2)$$

Here, the first term in the right-hand side of Eq. (2), $dV_{SET}/d\delta$, is E_{SET} on the basis of the definition $V_{SET} = V_{gap} + V_{sub} = E_{SET} \cdot \delta + \alpha \cdot d_{oxide}$ as explained in the main text. Next, the set time (t_{SET}) can be described as $t_{SET} = \delta/v$, where v is the drift velocity of V_O given as $v = a \cdot f \cdot \exp(-E_a/kT) \cdot \sinh(qaE_{SET}/2kT)$. Then, the second term in the right-hand side of Eq. (2), $d\delta/d\ln(t_{SET})$, can be derived as follows:

$$t_{SET} = \frac{\delta}{v} = \frac{\delta}{af\exp(-qE_a/kT)\sinh(qaE_{SET}/2kT)} = \frac{\delta}{af\exp(-qE_a/kT)\sinh(qaV_{gap}/2\delta kT)} \quad (3)$$

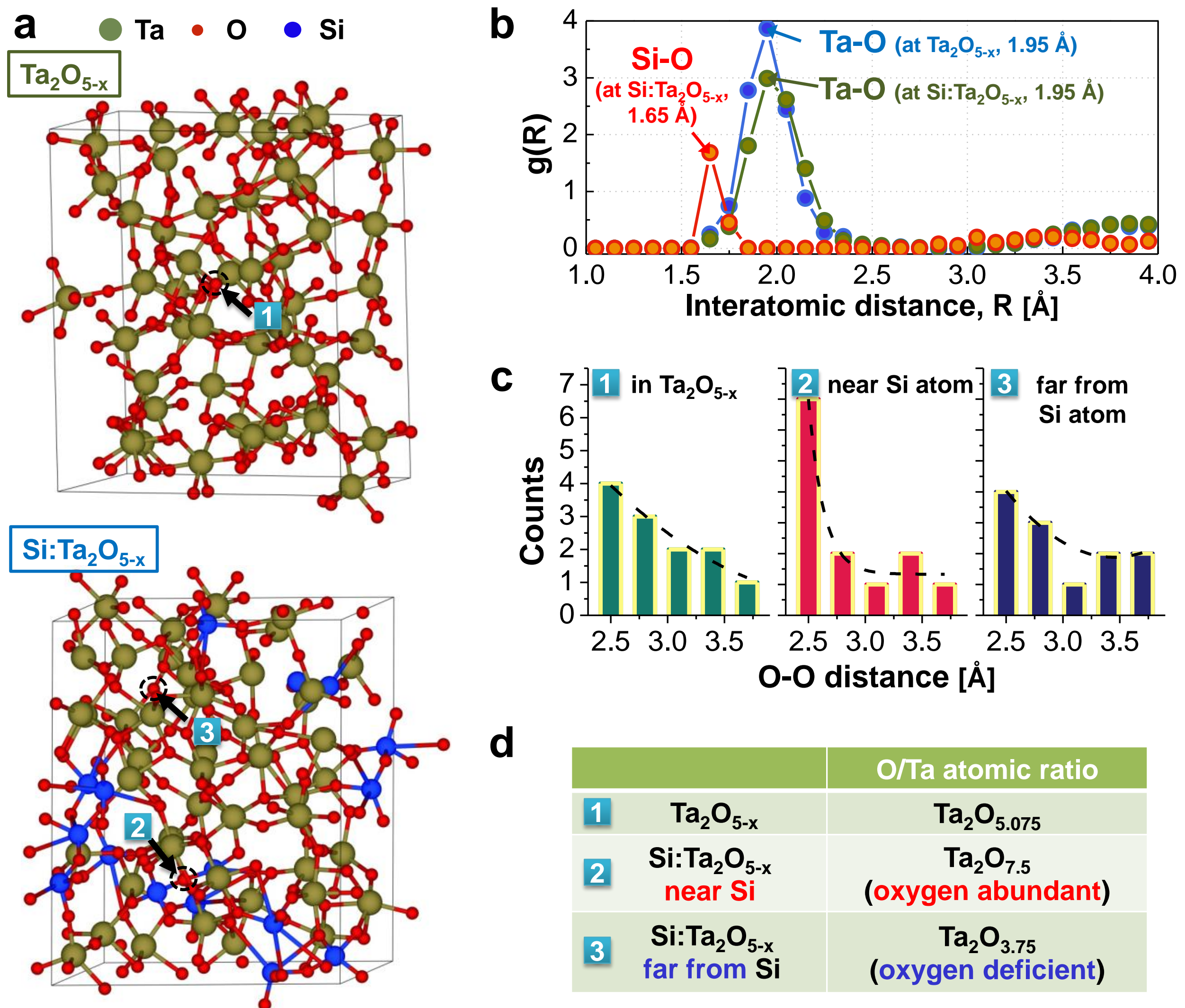
$$\frac{d\ln t_{SET}}{d\delta} = \frac{1}{\delta} + \frac{qaV_{gap}}{2\delta^2 kT}\left(\cosh(\frac{qaV_{gap}}{2\delta kT})/\sinh(\frac{qaV_{gap}}{2\delta kT})\right) \approx \frac{1}{\delta} + \frac{qaV_{gap}}{2\delta^2 kT} = \frac{1}{\delta}(1+\frac{qaE_{SET}}{2kT}) \approx \frac{qaE_{SET}}{2\delta kT} \quad (4).$$

Eq. 4 is obtained by noticing $\frac{qaV_{gap}}{2\delta kT} >> 1$ Finally, the slope of the V_{SET}–$\ln(t_{SET})$ curve is given as

$$slope = \frac{dV_{SET}}{d\delta} \cdot \frac{d\delta}{d\ln t_{SET}} = E_{SET} \cdot \left(\frac{2\delta kT}{qaE_{SET}}\right) = \left(\frac{2kT}{q}\right) \cdot \left(\frac{\delta}{a}\right). \quad (5)$$

5.7 ***Ab Initio*** Study

A question remained as to why the effective hopping distance is increased by the Si dopant. It has been believed that the hopping distance corresponds to the spacing between oxygen sites in the oxide [15]. To understand the nature of the hopping distance (and oxygen sites) modulation by dopant, we performed first-principle electronic-structure calculations on the



		O/Ta atomic ratio
1	Ta_2O_{5-x}	$Ta_2O_{5.075}$
2	$Si:Ta_2O_{5-x}$ near Si	$Ta_2O_{7.5}$ (oxygen abundant)
3	$Si:Ta_2O_{5-x}$ far from Si	$Ta_2O_{3.75}$ (oxygen deficient)

Figure 5.5. (a) Snapshots of the amorphous Ta_2O_5 and Si-doped Ta_2O_5 structures obtained in the *ab initio* simulation. The Ta, O, and Si atoms are colored in dark green, red, and blue, respectively. (b) Pair-correlation functions of the amorphous Ta_2O_5 and Si-doped Ta_2O_5 calculated at room temperature. (c) Histograms of the O–O distance from a selected oxygen atom to a neighboring oxygen atom. Three oxygen atoms are selected randomly, as shown in panel a. (d) O and Ta atomic ratio near the selected oxygen atoms.

basis of the density-functional theory using the Vienna *ab initio* simulation package (VASP). Figure 5.5(a) shows the calculated atomic structure of both undoped Ta_2O_5 and Si-doped Ta_2O_5 where a quenched amorphous Ta_2O_5 structure was used to better represent the experimental system, which is in contrast to previous *ab initio* studies based on crystal structures [16-18] Figure 5.5(b) shows the calculated pair correlation function, which represents the probability of

finding the center of a particle a given distance away from the center of another particle. In the undoped Ta_2O_5 case, the interatomic distance between Ta and O is 1.95 Å. When Si dopant is added, this Ta–O interatomic distance does not change. However, the interatomic Si–O distance is found to be shorter than the Ta–O distance, implying that oxygen can be located closer to Si than to Ta. In addition, the interatomic distances among oxygen atoms are investigated. Figure 5.4(c) shows the calculated O–O distance from the selected oxygen atom to a neighboring oxygen atom.

We select three arbitrary oxygen atoms, as shown in Figure 5.5(a): 1) an oxygen atom located in undoped Ta_2O_5 (case 1), 2) an oxygen atom located near a Si atom in Si:Ta_2O_5 (case 2), and 3) an oxygen atom located far away from the Si atoms in Si:Ta_2O_5 (case 3). The calculated O–O distance in Ta_2O_5 is centered around 0.3 nm, which is consistent with the measured hopping distance (0.32 nm, shown in Figure 3c). In addition, the O–O distance near Si (case 2) appears to be shorter than those of the other cases. From the data shown in Figures 5.5(b) and 5.5(c), we can conclude that the Si dopant more strongly attracts oxygen than Ta, and the oxygen atoms near Si are closely gathered. As a result, the region away from the Si dopant will turn into an oxygen-deficient state. This is confirmed by the *ab initio* calculations in Figure 5.5(d), which plots the calculated O and Ta atomic ratio in the three cases and indicates that an oxygen-deficient region is formed away from the Si dopant. The oxygen-deficient regions facilitate V_O transport in the Si doped devices as the V_Os can hop interstitially (*i.e.*, oxygen-deficient region away from Si dopant) or by substitution through the closely gathered oxygen atoms (*i.e.*, oxygen-abundant region near Si dopant). These processes allow effective ion hopping and explain the faster V_O drift under electric field in Si doped devices.

5.8 Analog Switching Behavior

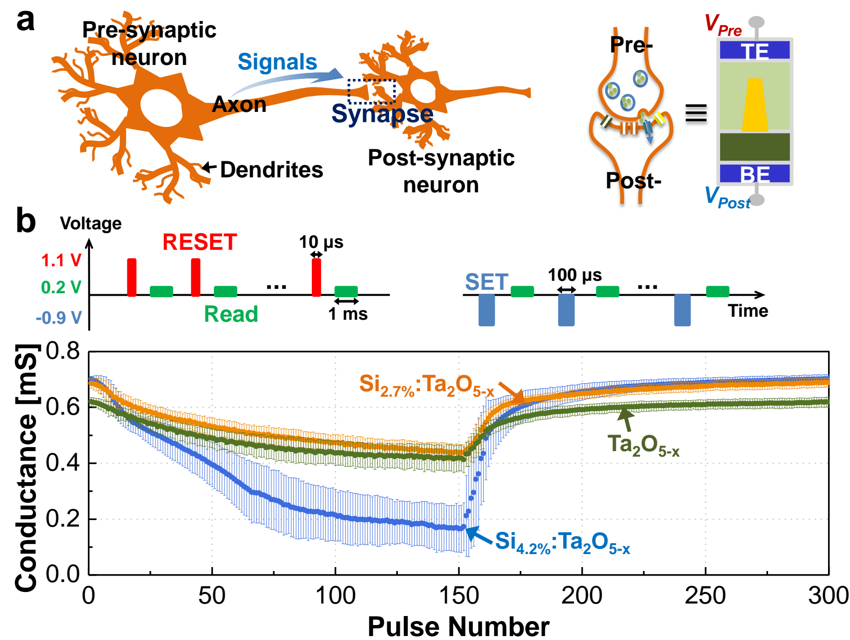


Figure 5.6. (a) Schematic illustration showing a synapse connecting a pair of neurons, where the synaptic functions can be emulated by RRAM devices. (b) Analog switching behavior obtained by pulse trains consisting of 150 reset pulses (1.1 V, 10 μs) followed by 150 set pulses (-0.9 V, 100 μs) with small, nonperturbative read voltage pulses (0.2 V, 1 ms) applied in the intervals. The conductance changes are measured during the read pulse and plotted as a function of applied pulse number. The error bars indicate the standard deviation from the measured data set, which are collected from 50 such test cycles in five different devices in each case.

Tantalum-oxide RRAM devices have been extensively studied due to its excellent endurance of over 10^{12}, which is the largest among all reported resistive devices [5]. However, the previously reported devices show mostly digital switching with limited dynamic range. Si doping improves the RS tunability and also leads to more incremental, analog-type conductance changes, making these devices also suitable for neuromorphic applications. In a neuromorphic system based on RRAM devices, the RRAM, whose weight can be incrementally modulated by electrical pulses ("spikes"), acts as a synapse connecting a pair neurons, as shown in Figure 5.6(a) [19]. Figure 5.6(b) shows the measured conductance values in the aforementioned three different

samples during the application of pulse (spike) trains. Each pulse train consists of 150 reset (depression) or set (potentiation) pulses, followed by read voltage pulses in the intervals. More intermediate states between maximum and minimum conductance in both the set and reset responses are clearly observed in the Si-doped devices as the number of applied pulses increases, offering a much larger dynamic range compared to the undoped devices. Further increasing the Si doping level results in higher dynamic range, but the switching becomes more digital-like due to the large hopping distance as shown in Figure 5.7 [9].

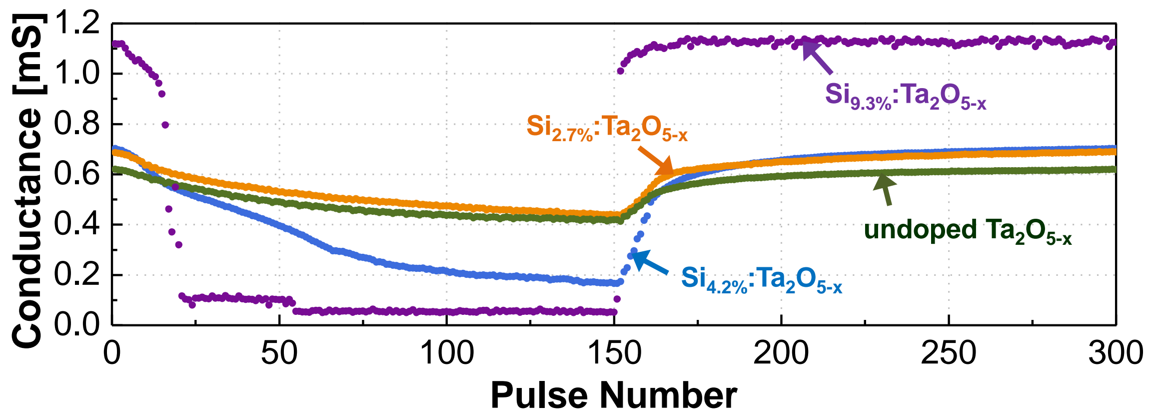


Figure 5.7. Analog switching behaviors obtained by pulse trains in four different cases.

Here, the more conductive LRS and consequently larger reset current in the Si-doped device may lead to higher power consumption during reset. However, this problem can be mitigated since Si doing also causes more resistive HRS and a larger dynamic range. As shown in Figures 5.6(b), although the maximum LRS conductance of the Si-doped device is higher than that of the undoped device, many intermediate conductance states with values lower than that of the undoped device can be obtained. The overall power consumption can be reduced if the devices are mostly cycled between these states, *e.g.*, as analog switches in neuromorphic applications. In addition, the conductance variation of the Si-doped devices appears larger than

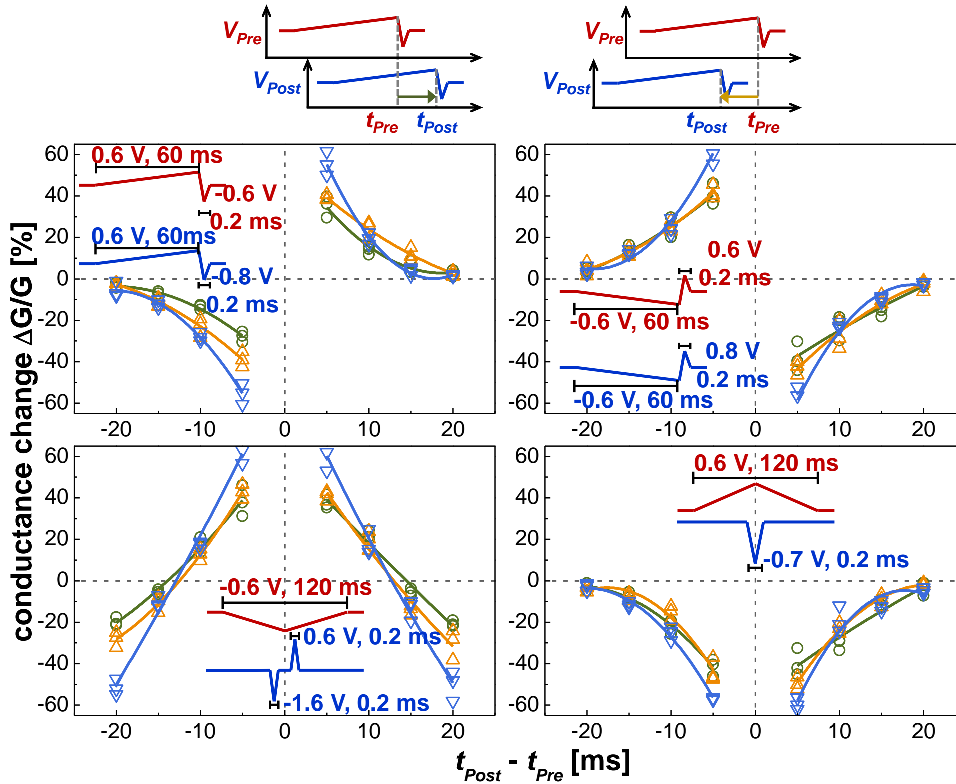


Figure 5.8. Implementing four different types of STDP using tantalum oxide RRAM. The pre-spike voltage (V_{Pre}) and post-spike voltage (V_{Post}) are applied to the TE and BE of the RRAM, respectively. The net programming voltage ($V_{Pre} - V_{Post}$) applied across the device depends on the positive or negative moments $t_{Post} - t_{Pre}$. The dots indicate the experimental data, and the lines are guides to the eye. The insets show the (red) pre- and (blue) postsynaptic spike schemes.

the undoped devices. We hypothesize that this larger variation may be caused by the larger hopping distances due to Si doping. With larger hopping distance, the stochastic properties of ion hopping will be more pronounced since the stochastic movement of just a few ions may already cause significant resistance changes which in turn affect the dynamic CF growth and dissolution processes, causing variations in the CF shape and resistance variations. On the contrary, small hopping distances mean many ions need to be moved to cause significant resistance change so the stochastic hopping properties of individual ions are more effectively averaged out, leading to smaller conductance variations.

The ability of the Si-doped tantalum-oxide devices was further tested by implementing Spike-timing-dependent-plasticity (STDP) learning rules [20,21]. STDP refers to the effect that the relative timing of pre- and postsynaptic spikes determines the sign and magnitude of the long-term synaptic weight change, which can potentially appear in four different forms [22-24]. These four different STDP forms are successfully implemented in our tantalum-oxide-based RRAM devices by designing spike-pairing protocols, as shown in Figure 5.8. Clearly, the extent of conductance change depends on the amount of Si dopant; and improved analog performance is obtained through Si-doped tantalum RRAM devices.

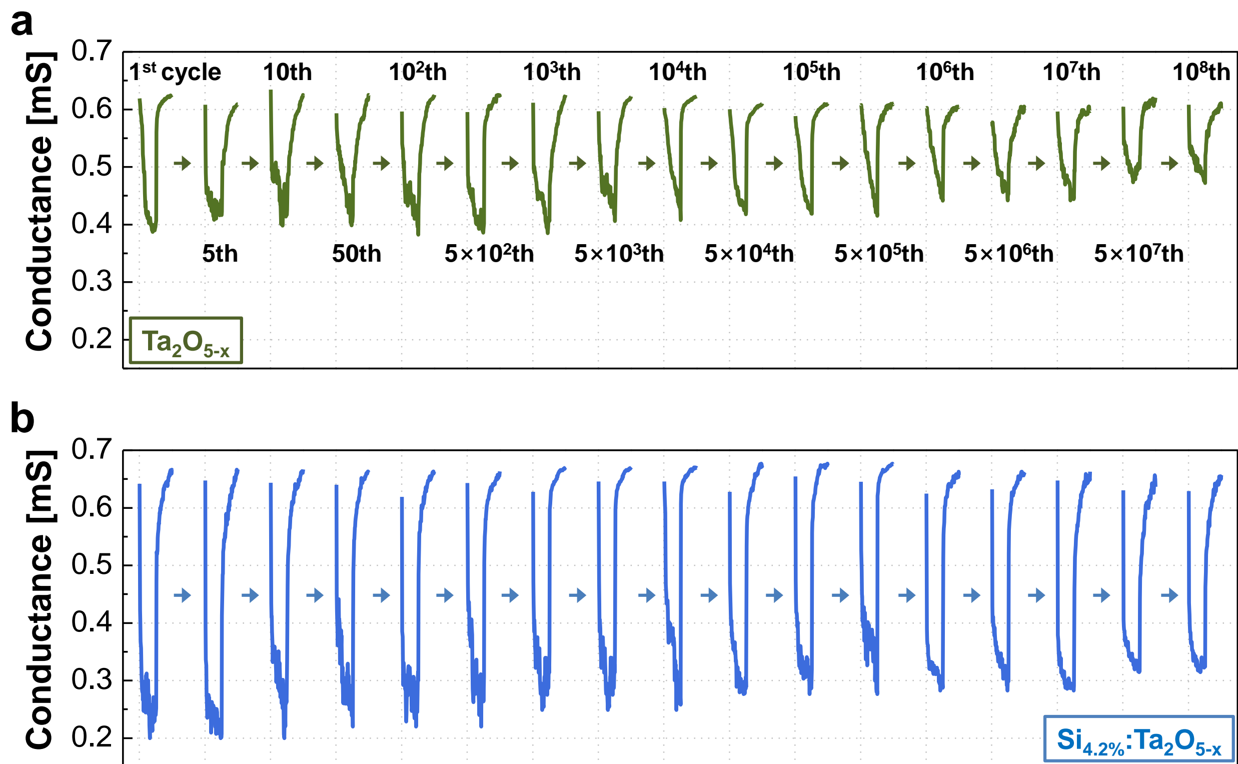


Figure 5.9. Measured cycling endurance performance of analog switching in (a) a Ta_2O_{5-x} RRAM and (b) a $Si_{4.2\%}$:Ta_2O_{5-x} RRAM. Each test cycle consists of a pulse train including 50 reset (1.25 V, 10 µs) pulses followed by 50 set (-1.0 V, 10 µs) pulses.

Finally, we show that the excellent cycling property of the tantalum oxide RRAM devices is preserved after Si doping, as shown in Figure 5.9. The 4.2% Si doped device still maintains analog RS behavior over 10^8 test cycles, with each test cycle containing 50 reset and 50 set pulses corresponding to over 10^{10} total set/reset operations. This reliable analog RS behavior ensures stable long-term operation and will help the development of large-scale RRAM-based neuromorphic systems with designable synaptic functions.

5.9 Conclusion

In conclusion, we show that the RS behavior in RRAM devices can be systematically tuned at the atomic level through doping. Specifically, Si doping can cause faster V_O drift and improves the RS characteristics and leads to more controllable analog switching behavior. A measurement methodology was developed to extract the hopping distance and the depleted gap length during V_O migration. The experimental findings were supported by *ab initio* calculations. We believe these results not only produce a desired RRAM system that can be directly used in neuromorphic computing applications, but also provide guidance for continued design and optimization this important class of devices.

5.10 References

[1] Ielmini, D. Modeling the Universal Set/Reset Characteristics of Bipolar RRAM by Field- and Temperature-Driven Filament Growth. IEEE Trans. Electron Dev. 2011, 58, 4309-4317.

[2] Strukov, D. B.; Snider, G. S.; Stewart, D. R.; Williams, R. S. The Missing Memristor Found. Nature 2008, 453, 80-83.

[3] Yang, J. J.; Strukov, D. B.; Stewart, D. R. Memristive Devices for Computing. Nat. Nanotechnol. 2013, 8, 13-24.

[4] Yang, J. J.; Zhang, M. X.; Strachan, J. P.; Miao, F.; Pickett, M. D.; Kelley, R. D.; Medeiros-Ribeiro, G.; Williams, R. S. High Switching Endurance in TaOx Memristive Devices. Appl. Phys. Lett. 2010, 97, 232102.

[5] Lee, M. -J.; Lee, C. B.; Lee, D.; Lee, S. R.; Chang, M.; Hur, J. H.; Kim, Y. –B.; Kim, C. –J.; Seo, D. H.; Seo, S.; et al. A Fast, High-Endurance and Scalable Non-Volatile Memory Device Made from Asymmetric Ta2O5-x/TaO2-x Bilayer Structures. Nat. Mater. 2011, 10, 625-630.

[6] Sadaf, S. M.; Liu, X.; Son, M.; Park, S.; Choudhury, S. H.; Cha, E.; Siddik, M.; Shin, J.; Hwang, H. Highly Uniform and Reliable Resistance Switching Properties in Bilayer WOx/NbOx RRAM Devices. Phys. Status Solidi A 2012, 209, 1179-1183.

[7] Yang, Y.; Choi, S.; Lu, W. Oxide Heterostructure Resistive Memory. Nano Lett. 2013, 13, 2908-2915.

[8] Fang, Z.; Yu, H. Y.; Li, X.; Singh, N.; Lo, G. Q.; Kwong, D. L. HfOx/TiOx/HfOx/TiOx Multilayer-Based Forming-Free RRAM Devices With Excellent Uniformity. IEEE Elect. Dev. Lett. 2011, 32, 566-568.

[9] Kim, S.; Choi, S. H.; Lu, W. Comprehensive Physical Model of Dynamic Resistive Switching in An Oxide Memristor. ACS Nano 2014, 8, 2369-2376.

[10] Strukov, D. B.; Williams, R. S. Exponential Ionic Drift: Fast Switching and Low Volatility of Thin-Film Memristors. Appl. Phys. A 2009, 94, 515-519.

[11] Fromhold, A. T.; Cook, E. L. Diffusion Currents in Large Electric Fields for Discrete Lattices. J. Appl. Phys. 1967, 38, 1546-1553.

[12] Jo, S. H.; Kim, K. H.; Lu, W. Programmable Resistance Switching in Nanoscale Two-Terminal Devices. Nano Lett. 2009, 9, 496-500.

[13] Ambrogio, S.; Balatti, S.; Gilmer, D. C.; Ielmini, D. Analytical Modeling of Oxide-Based Bipolar Resistive Memories and Complementary Resistive Switches. IEEE Trans. Elect. Dev. 2014, 61, 2378-2385.

[14] Miao, F.; Yang, J. J.; Strachan, J. P.; Stewart, D.; Williams, R. S.; Lau, C. N. Force Modulation of Tunnel Gaps in Metal Oxide Memristive Nanoswitches. Appl. Phys. Lett. 2009, 95, 113503.

[15] Goux, L.; Sankaran, K.; Kar, G.; Jossart, N.; Opsomer, K.; Degraeve, R.; Pourtois, G.; Rignanese, G. M.; Detavernier, C.; Clima, S.; et al. Field-Driven Ultrafast Sub-ns Programming in W/Al2O3/Ti/CuTe-based 1T1R CBRAM System. 2012 Symposium on VLSI Technology, Hawaii, USA, Jun. 12-14 2012; 69-70.

[16] Zhao, L.; Ryu, S. –W.; Hazeghi, A.; Duncan, D.; Magyari-Kope, B.; Nishi, Y. Dopant Selection Rules for Extrinsic Tunability of HfOx RRAM Characteristics: A Systematic Study. 2013 Symposium on VLSI Technology, Kyoto, Japan, Jun. 11-13, 2013; 106-107.

[17] Zhang, H.; Liu, L.; Gao, B.; Qiu, Y.; Liu, X.; Lu, J.; Han, R.; Kang, J.; Yu, B. Gd-Doping Effect on Performance of HfO2 Based Resistive Switching Memory Devices Using Implantation Approach. Appl. Phys. Lett. 2011, 98, 042105.

[18] Zhang, H.; Gao, B.; Sun, B.; Chen, G.; Zeng, L.; Liu, L.; Liu, X.; Lu, J.; Han, R.; Kang, J.; et al. Ionic Doping Effect in ZrO2 Resistive Switching Memory. Appl. Phys. Lett. 2010, 96, 123502.

[19] Bear, M. F.; Malenka, R. C. Synaptic Plasticity: LTP and LTD. Curr. Opin. Neurobiol. 1994, 4, 389-399.

[20] Jo, S. H.; Chang, T.; Ebong, I.; Bhadviya, B. B.; Mazumder, P.; Lu, W. Nanoscale Memristor Device as Synapse in Neuromorphic Systems. Nano Lett. 2010, 10, 1297-1301.

[21] Yu, S.; Wu, Y.; Jeyasingh, R.; Kuzum, D.; Wong, H. –S. P. An Electronic Synapse Device Based on Metal Oxide Resistive Switching Memory for Neuromorphic Computation. IEEE Trans. Elect. Dev. 2011, 58, 2729-2737.

[22] Abbott, L. F.; Nelson, S. B. Synaptic Plasticity: Taming the Beast. Nat. Neurosci. 2000, 3, 1178-1183.

[23] Roberts, P. D.; Bell, C. C. Spike Timing Dependent Synaptic Plasticity in Biological Systems. Biol. Cybern. 2002, 87, 392-403.

[24] Li, Y.; Zhong, Y.; Xu, L.; Zhang, J.; Xu, X.; Sun, H.; Miao, X. Ultrafast Synaptic Events in A Chalcogenide Memristor. Sci. Rep. 2013, 3, 1619.

Chapter 6.
Data Clustering using RRAM network

6.1 Introduction

In the previous chapters, we investigated the properties of resistive switching and discussed how to improve analog switching behavior of RRAM. In this chapter, we demonstrate a potentially important application of RRAM networks - data clustering based on unsupervised learning.

The von Neumann architecture, widely used in conventional computing systems, has become less optimal in data-intensive tasks due to limited data transfer rates between the memory and the central processing unit. Alternative computing systems such as neuromorphic or machine learning systems, have attracted increasing attention in dealing with "big data" problems such as pattern recognition from large amounts of data sets [1, 2]. Principal component analysis [3] is an important technique used in machine learning to discover orthogonal factors underlying multivariate data by examining the correlations among the set of input variables. The technique can also be used to reduce the dimensionality of input data and cluster data for subsequent data classification, and is thus an important preprocessing step for many machine learning algorithms. Here we show that principal component analysis (PCA) can be efficiently achieved in simple RRAM-based crossbar networks with online learning capability, allowing this technique to be used to effectively classify sensory data.

The two key factors that make RRAM crossbar arrays attractive for neuromorphic or machine learning systems are 1) their ability to naturally implement matrix operations (e.g. dot-product): due to the resistive nature of the two-terminal device, the RRAM crossbar array can directly convert an input voltage vector into an output current (or charge) vector, weighed by the RRAM conductance at each matrix element, thus directly and efficiently performing the matrix operation; and 2) their ability to achieve online learning with simple programming pulses: the weights of the RRAM crossbar matrix - the device conductances, can be incrementally trained using simple voltage pulses [4-5]. Other properties such as high density, low power consumption, long cycling endurance and subnanosecond switching speed have also been demonstrated in RRAM devices [6–10].

A typical RRAM device consists of a transition metal oxide layer such as TiOx, HfOx, WOx sandwiched by a pair of electrodes [11-13]. The resistance of the RRAM device can be adjusted by controlling the amount and distribution of oxygen vacancies, which modulate the local conductivity in the oxide layer [14, 15]. Using an unsupervised, online learning rule, we show that crossbar arrays of RRAM devices can learn the principal components from sensory data (e.g. database of breast cancer measurements) and effectively separate unlabeled data into clusters. After data clustering, a conventional supervised learning process can then be used to define a decision boundary and effectively classify tumors as malignant or benign.

6.2 Device Fabrication and Measurement Setup

The device fabrication starts with a highly p-doped Si/SiO_2 substrate with a 100 nm thermal SiO_2 layer. The bottom electrodes (BEs) were fabricated by photolithography and liffoff

with 5nm-thick NiCr and 40nm-thick Pd. The 40 nm of oxygen-rich TaO_x layer was sputtered by direct current (DC) reactive using a Ta metal target with Ar(97%)/O_2(3%) gas mixture at 400 ℃. Next, 5 nm of Ta_2O_5 switching layer was sputtered by 140W radio frequency (RF) sputtering while p-doped Si was co-sputtered with Ta_2O_5 layer with 70W DC sputtering at room temperature. The top electrodes (TEs) with 40nm of Pd and 20nm of Au were fabricated by photolithography and liffoff to form a crossbar structure. The electrical characterization were performed with a custom-built measurement system in a probe station (Desert Cryogenics TTP4).

6.3 Analog RRAM Behavior

The analog switching behavior is obtained from a tantalum-oxide RRAM based on a bilayer structure consisting of an oxygen-rich Ta_2O_5 layer and an oxygen-deficient TaO_x layer [6,10,14,16]. We have shown that such a RRAM with the tantalum oxide layer doped with silicon atoms can show improved dynamic range and controllable analog switching behavior [17]. In this study, 2 μm × 2 μm devices and crossbar arrays were used following the processes discussed in Ref. [17]. During measurements, the bias voltage was applied to the top electrode (TE) while the bottom electrode (BE) was grounded. Fig. 6.1(a) shows DC current – voltage (I-V) curve of a device showing typical bipolar resistive switching characteristics. In this system, an applied voltage can change the amount and distribution of oxygen vacancies and modulate the conductive channels in the Ta_2O_5 layer which controls the conductance of the device [14-17], as schematically shown in Fig. 6.1(b).

To model the conductance change of the RRAM, we introduce the internal state variable, *w*, which serves as an area index representing the number of conductive filaments or,

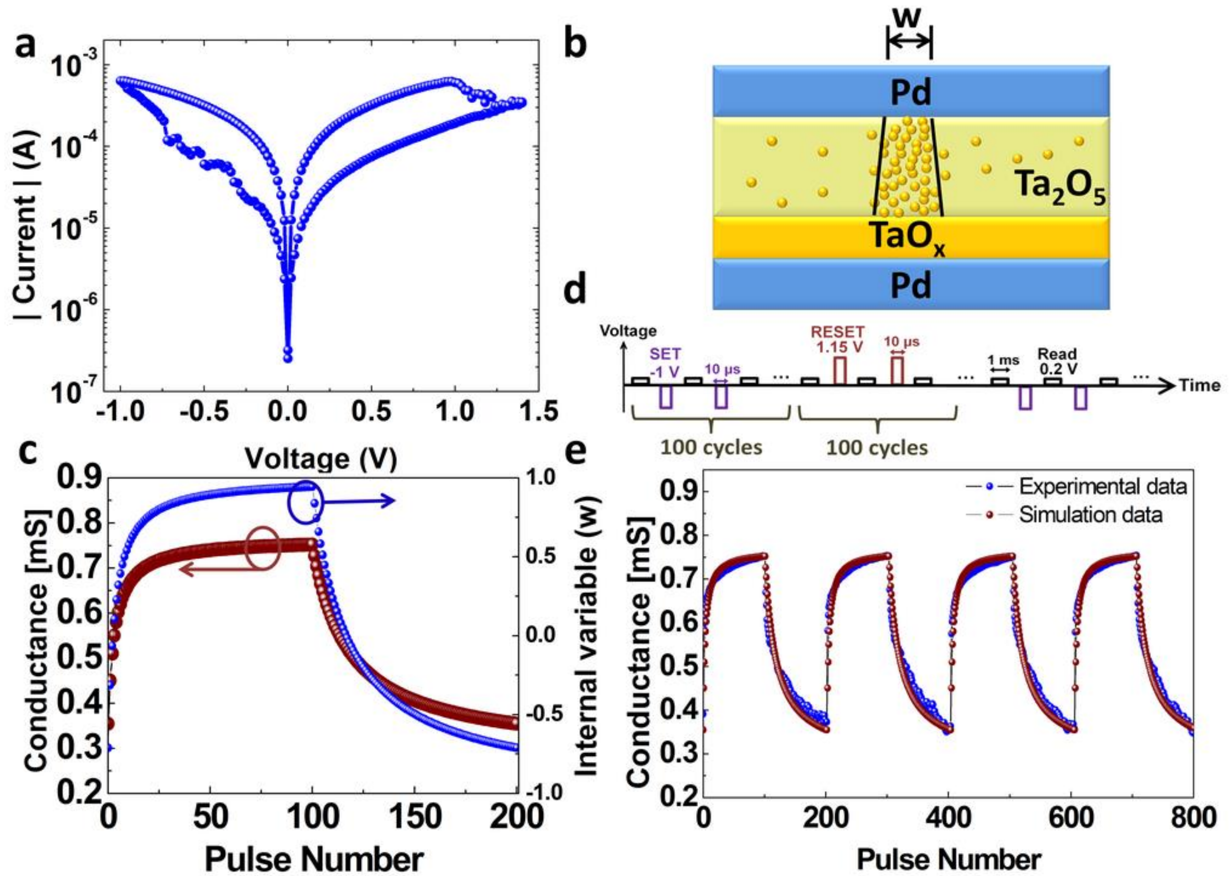


Figure 6.1. Modelling the switching performance of a RRAM. (a) DC I-V characteristics of a typical RRAM device showing the bipolar switching. (b) Schematic image of a RRAM device having oxygen vacancy filament. (c) Calculated conductance and internal state variable with 100 pulses of potentiation (-1 V, 10μs) and depression (1.15 V, 10μs), consecutively. (d) The sequences of the applied pulses showing 4 sets of 100 pulses of potentiation and 100 pulses of depression. (e) The measured and calculated conductance changes measured by read (0.2V) pulse with the set and reset processes shown in Fig. 1(d).

equivalently, the area covered by the conductive channel as shown in Fig. 6.1(b). The dynamics of the state variable in response to the applied voltage is described by equation (1), where *u*() is

the Heaviside step function, k, μ_1, u_2, are positive parameters determined by material properties such as ion hopping distance and hopping barrier heights [13].

$$\frac{dw}{dt} = (w-1)^2 k(e^{-\mu_1 V} - e^{\mu_2 V})u(-V) + w^2 k(e^{-\mu_1 V} - e^{\mu_2 V})u(V) \quad (1)$$

$$I = w\,\gamma \sinh(\delta \times V) + (1-w)\alpha\left(1 - e^{-\beta \times V)}\right) \quad (2)$$

The current through the device is described by equation (2) which consists of the term describing conduction through the channel area (tunneling-dominated conduction, first term) and the rest of the device (Schottky-dominated conduction, second term) [13]. This equation clearly shows how the device conductance is regulated by the state variable, w. γ, δ, α, β are positive parameters determined by material properties such as the effective tunneling distance, tunneling barrier, the depletion width of the Schottky barrier region and Schottky barrier height [13] (Appendix 6A). The RRAM model, consisting of the state variable dynamic equation (1) and *I-V* equation (2), was tested against experimental measurements. For example, in Fig. 1c, pulse programming conditions were simulated with the application of a train of one-hundred -1 V, 10 μs pulses followed by a train of one-hundred 1.15 V, 10 μs pulses, with the device conductance monitored with a 0.2 V read pulse after each training pulse. With the application of a negative pulse, the RRAM conductance gradually increases (purple curve), followed by the increase in the internal state variable value (blue curve). On the other hand, a positive pulse decreases the conductance following the decrease of the internal state variable value. The experimental data measured form an actual RRAM device and the simulation data were compared and plotted together in Fig. 1d, showing that the model can trace the experimental data precisely.

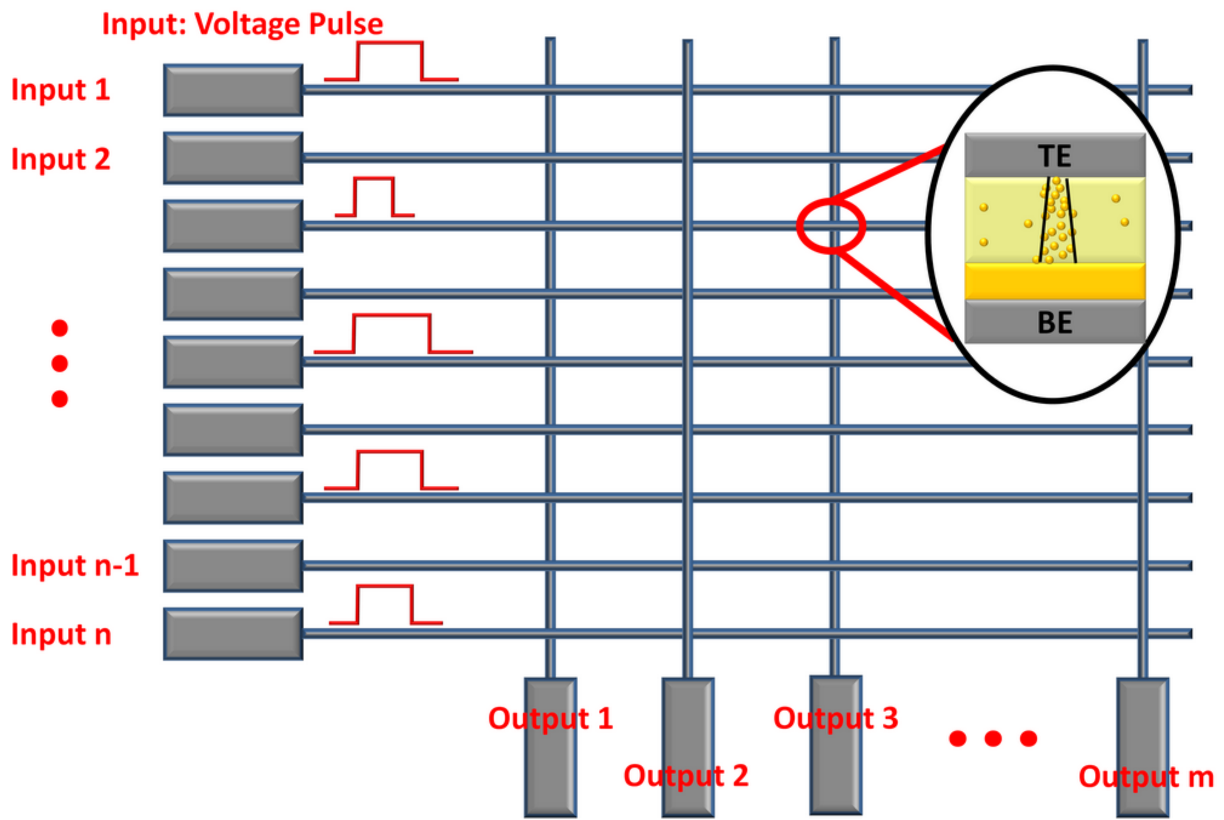


Figure 6.2. The network schematic. The column electrodes represent inputs and the row columns represent outputs. The RRAM devices are located at the intersections where the column electrodes and row electrodes connected.

6.4 **Learning in Crossbar Arrays**

To implement PCA, we adopted a neural network structure using a crossbar array of RRAMs as shown in Fig. 6.2, where the *n* input channels are connected to the rows and the *m* output channels are connected to the columns of the RRAM crossbar network. In this study, a standard breast cancer data set from University of Wisconsin Hospitals, Madison was used as the input signal data [18, 19]. The data set consists of breast cell mass properties in 9 categories including clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. The sensory data were derived from a digitized image of a fine needle aspirate (FNA) of a breast mass and

each category has a range from 0 to 10. In a feature learning test, the measurement results from the 9 categories of a given cell are fed to the 9 inputs (n = 9) of the neural network, and the output is obtained from the 2 output channels (m = 2). The input signals are implemented as voltage pulses with fixed amplitude (0.2 V) and variable pulse widths proportional to the measured values in the corresponding category. Each training cycle consists of one hundred randomly sequenced data points (50 points from benign class, 50 points from malignant class). Afterwards, the ability of the network to successfully cluster the data and classify a cell as either benign or malignant was tested using 583 data points (not included in the training set).

As discussed earlier, in this configuration the output vector is determined by the dot-product of the input vector and the RRAM weight matrix. Additionally, the network learns the principal components by adjusting the RRAM weights during training. In this study, starting from a RRAM network with randomly distributed weights, we employ Sanger's rule (also known as the generalized Hebbian algorithm) to implement online learning to learn the principal components of the input data set. Sanger's rule is derived from Hebb's learning rule [20, 21] and these model learning rules have been widely adapted in artificial neural networks. Specifically, Sanger's rule utilizes the weight (g), output response (y) and present input (x) as shown in equation (3).

$$\Delta g_{ij} = \eta y_j (x_i - \sum_{k=1}^{j} g_{ik} y_k) \tag{3}$$

where η is the learning rate and is typically a small positive value($\ll 1$), x_i represents the input pulse at input (row) i and the value of the data is represented by the pulse width, and j = 1 or 2

corresponds to the primary principal component and the second principal component, respectively. g_{ij} is the weight at row i and column j in the network. Specifically, g_{ij} is defined as

$$g_{ij} = 2w_{ij} - 1 \tag{4}$$

where w_{ij} is the state variable of the RRAM device at row i and column j as discussed in Eq. (1). While w is positive only g_{ij} ranges from -1 to 1 from the definition. Note no label is used in the learning process. After training, the weights in columns 1 and 2 form the (first and 2nd, respectively) principal components of the input data set [21]. Accordingly, outputs obtained from the trained network will be clustered and can be used in subsequent classification analysis.

Specifically, with the application of an input x_j, the amount of charge collected at the output in the RRAM network can be obtained as:

$$Q_j = \sum_i \left[w_{ij}A + \left(1 - w_{ij}\right)B\right] x_i \tag{5}$$

where the charge is assumed to be determined by the current (Eq. (2)) and linearly proportional to the applied pulse width (x_i), and the constants in Eq. (2) have been lumped into constants A and B. The output, y_j, is then obtained from the charge Q_j through the following equation (6).

$$y_j = \frac{2Q_j}{A-B} - \sum_i \left[\frac{A+B}{A-B} x_i\right] \tag{6}$$

Plugging Eqs. (4)-(5), (6) can be simplified as:

$$y_j = \sum_{i=1}^{n} g_{ij} x_i \tag{7}$$

As expected, by properly choosing the output definition (here linearly dependent on the charge, eq. 6), the obtained output *y* corresponds to the vector product of the input and the weight matrix, as required by neural network algorithms.

During the training phase, the output is first obtained (by applying a 0.2 V read voltage with a pulse width proportional to the value of the training data at each column) from the RRAM array using equation (6), and the desired weight update Δg_{ij} is then calculated based on equation (3). Programming voltage pulses are then applied to the inputs to modify the RRAM weights. The programming pulses are determined by the polarity and magnitude of Δg_{ij}, with potentiation (-1 V) pulses applied to the input for positive Δg_{ij} and depression (1.15 V) pulses for negative Δg_{ij}, while the pulse widths are determined by the magnitude of $|\Delta g_{ij}|$. To account for the non-linear response of *w* with respect to training pulse (*i.e.* the effectiveness of weight change *dw*/*dt* depends on the device state *w*, as evidenced in Eq. 1 and Fig. 6.1 (c)-(d)), a compensation scheme is employed to ensure the desired conductance change. Specifically, the pulse width $|\Delta t|$ is determined as

$$\Delta t_{ij} = \frac{2}{k\left(e^{-\mu_1 V_{potentiation}} - e^{\mu_2 V_{potentiation}}\right)}\left(\frac{-1}{g_{ij,after}-1} + \frac{1}{g_{ij,before}-1}\right) u(\Delta g)$$

$$+ \frac{2}{k\left(e^{-\mu_1 V_{depression}} - e^{\mu_2 V_{depression}}\right)}\left(\frac{-1}{g_{ij,after}+1} + \frac{1}{g_{ij,before}+1}\right) u(-\Delta g) \quad (8)$$

When applied to equation (1) and by noticing the relationship between *w* and *g* ($g = 2w - 1$), equation (7) leads to the desired weight change in equation (3).

6.5 Details of the Training Process

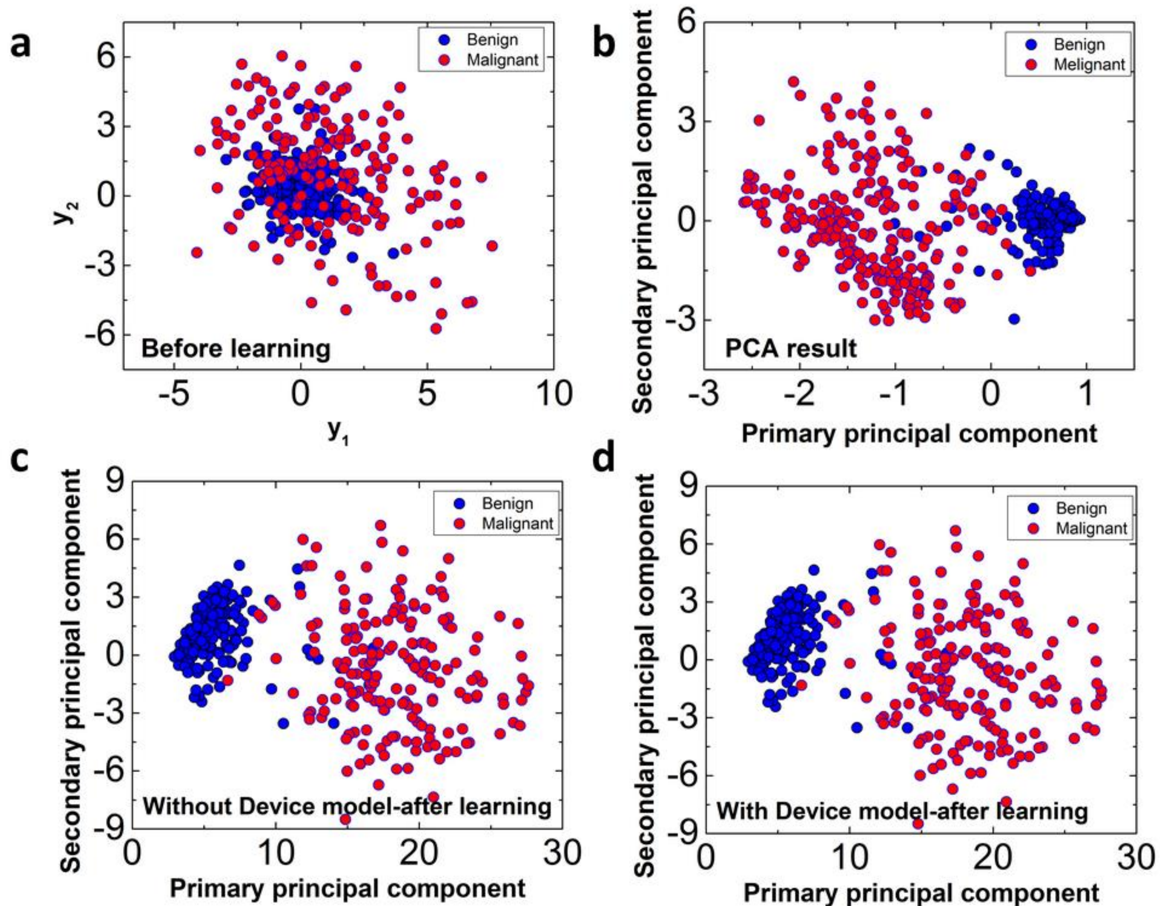


Figure 6.3. The result of principal component analysis. (a) The result of read process through RRAM devices showing y_1 at horizontal axis and y_2 at vertical axis before learning process. (b) Principal component analysis using traditional covariance matrix of the data. (c) Principal component analysis using Sanger's rule without the RRAM model. (d) Principal component analysis using Sanger's rule with the RRAM model.

Fig. 6.3(a) shows results of the 583 test data points before learning (*e.g.* when the RRAM weights are random), with y_1 at horizontal axis and y_2 at vertical axis. Blue dots and purple dots represent benign and malignant cells (the ground truth), respectively. We note the labels were not used during training and are only shown here to illustrate the effectiveness of the

clustering process. It's clear from Fig. 6.3(a) that before training the benign set and the malignant set significantly overlap each other. In other words, the network before learning cannot effectively cluster the sets (with untrained, random weights). Results obtained after performing classical PCA calculations by directly calculating the eigenvectors and eigenvalues of the data covariance using matrix operations are shown in Fig. 6.3(b). The PCA calculations perform orthogonal transformation to identify the primary principal component in the direction of the largest variance, and subsequently the 2nd principal component, etc [22]. As expected, the data become clustered after transforming the data along the first two principal components, as shown in Fig. 6.3(b).

Instead of directly calculating the principal components using matrix operations and existing data, the principal components can also be obtained through training in neural networks, as discussed earlier. Fig. 6.3(c) shows results obtained from an idealized neural network using Sanger's rule, using only equation (3) and equation (7) without considering the physical RRAM device model. Successful linear separation of the data sets was also achieved in the neural network. In this case, instead of computed from current data set, the principal components were learned using Sanger's rule and are represented by the weights associated with specific outputs. More importantly, Fig. 6.3(d) shows the results obtained in the neural network employing the physical RRAM device model during training and feature extraction analysis. Successful clustering of the data, similar to the ones obtained from direct PCA calculations and learning with an ideal neural work, was also obtained in the RRAM network, suggesting the potential of the RRAM networks for feature learning tasks with online, unsupervised learning.

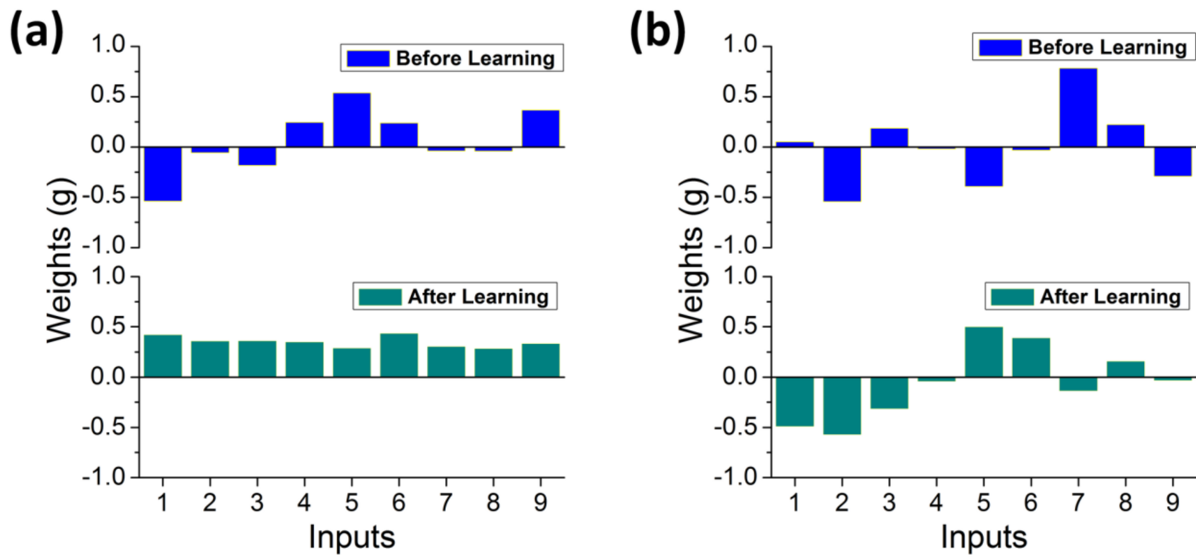


Figure 6.4. Weights distribution changes for (a) the primary principal component, (b) the second principal component before and after learning process.

Fig. 6.4 shows the primary and secondary principal components learned in the RRAM network from the training process, represented by the two 9-dimentional weight vectors associated with the two outputs. The training consists of 1000 training cycles. Since the application of Sanger's rule automatically normalizes the weights the Euclidean norm of the

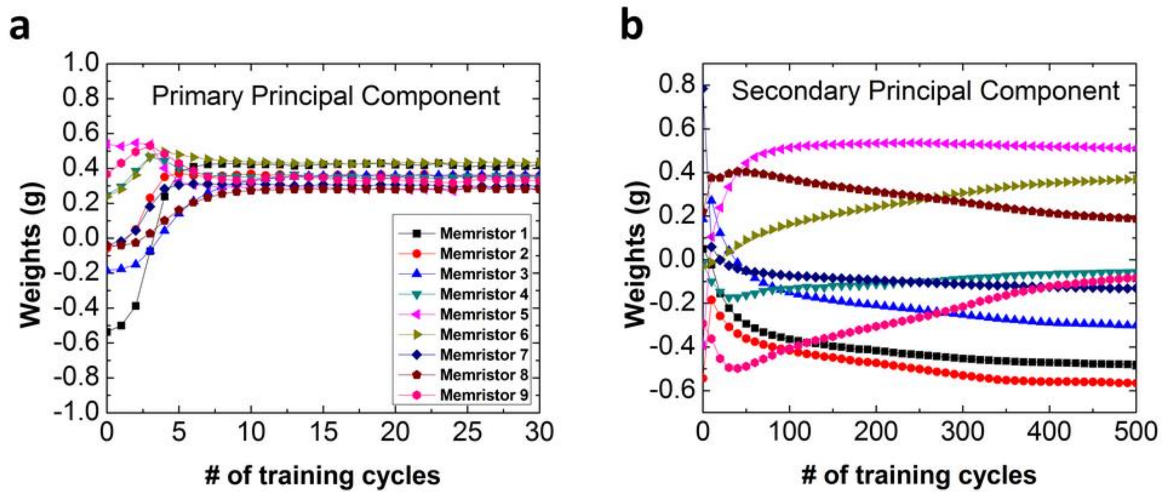


Figure 6.5. Weights changes with individual learning cycles for (a) the primary principal component, (b) the secondary principal component.

weights should converge to 1 after training (Appendix 6B). Indeed, the length of the weight vector for the primary principal component was found to converge from 0.9 to 1.0005 and that for the secondary principal component was found to converge from 1.12 to 1.003. In practice, this normalization condition can be used to determine when the network has completed learning.

To examine how the weights change during learning, weight distributions for the first two principal components during training are plotted in Fig. 6.5. For the primary principal component (Fig. 6.5(a)), the weights change rapidly in the first 10 cycles and quickly become stabilized for the rest of the learning cycles. While for the secondary principal component (Fig. 6.5(b)) the weights change gradually and the distribution stabilizes at a much later time. The reason for the different behaviors lie in the fact that for the primary principal component, only y_1 and g_{i1} need to be taken into account during weight update (equation (3)); however, for the secondary principal component, both y_1, y_2, and g_{i1} and g_{i2} need to be considered so convergence of the secondary principal component is more difficult and only happens after the primary principal component has stabilized.

6.6 **Effect of Applied Voltage Amplitude and the Learning Rate**

The effect of the applied voltage during learning and the learning rate are shown in Fig. 6.6. Fig 6.6(a) shows the histogram graphs of the number of pulses used during the training processes for different pulse amplitudes, measured in 20ns intervals. As expected, it can be seen that lower potentiation/depression voltages requires longer pulse widths in general, while faster learning can be obtained at higher voltages. Additionally, Fig. 6.6(b) shows the effect of the learning rate, η, on the training process. The weight redistribution for the secondary principal

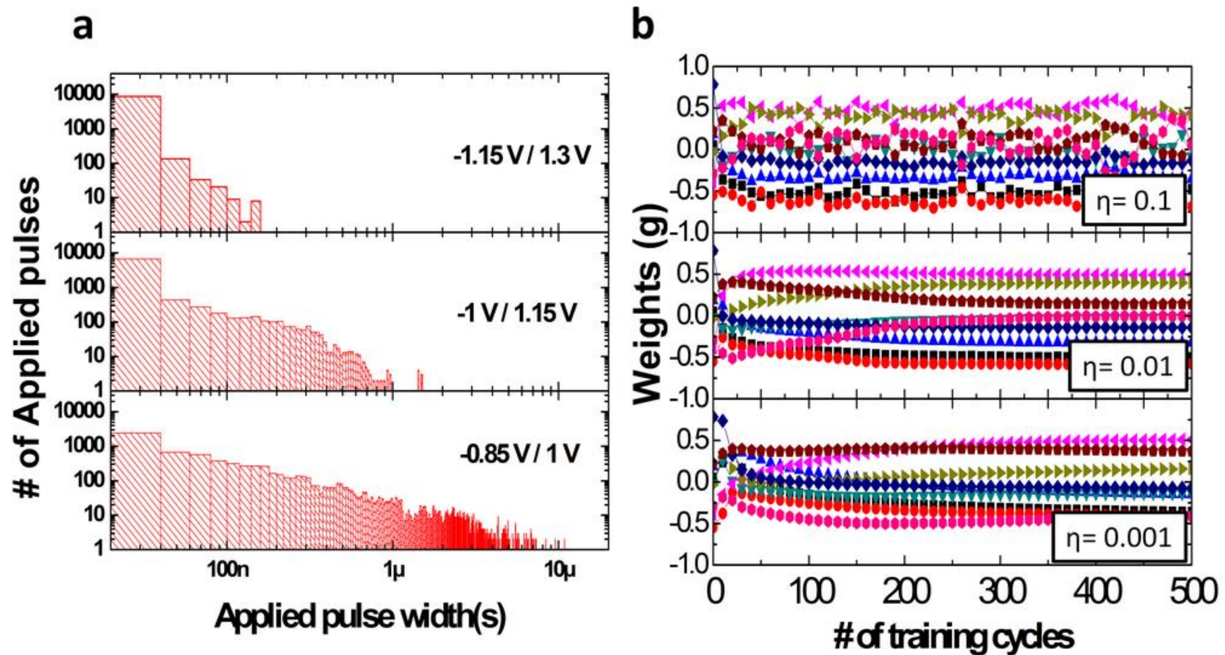


Figure 6.6. The effects of potentiation/depression voltage amplitudes and learning rate changes. (a) The histogram of applied pulse widths as a function of potentiation/depression voltage amplitude. (b) The weight changes as a function of learning rate.

component as a function of training is plotted. If the learning rate is too high (η=0.1), weight update becomes too fast (Eq. 3) and can overshoot the optimal value. As a result, the weight distributions fluctuate during training and never fully stabilize, as shown in the top graph in Fig. 6b. On the other hand, if the learning rate is too small (η=0.001), the weight updates becomes very slow and may not be able to overcome local minima, as shown in the bottom graph in Fig. 6.6(b). A properly chosen learning rate (η=0.01) balances learning speed and accuracy.

6.7 The Effect of Device Non-Uniformity

In the following, we discuss the effects of device-device variations in the network performance. Nanoscale devices such as RRAMs whose operations are essentially based on defects (e.g. oxygen vacancies) are intrinsically less reliable than conventional transistor devices.

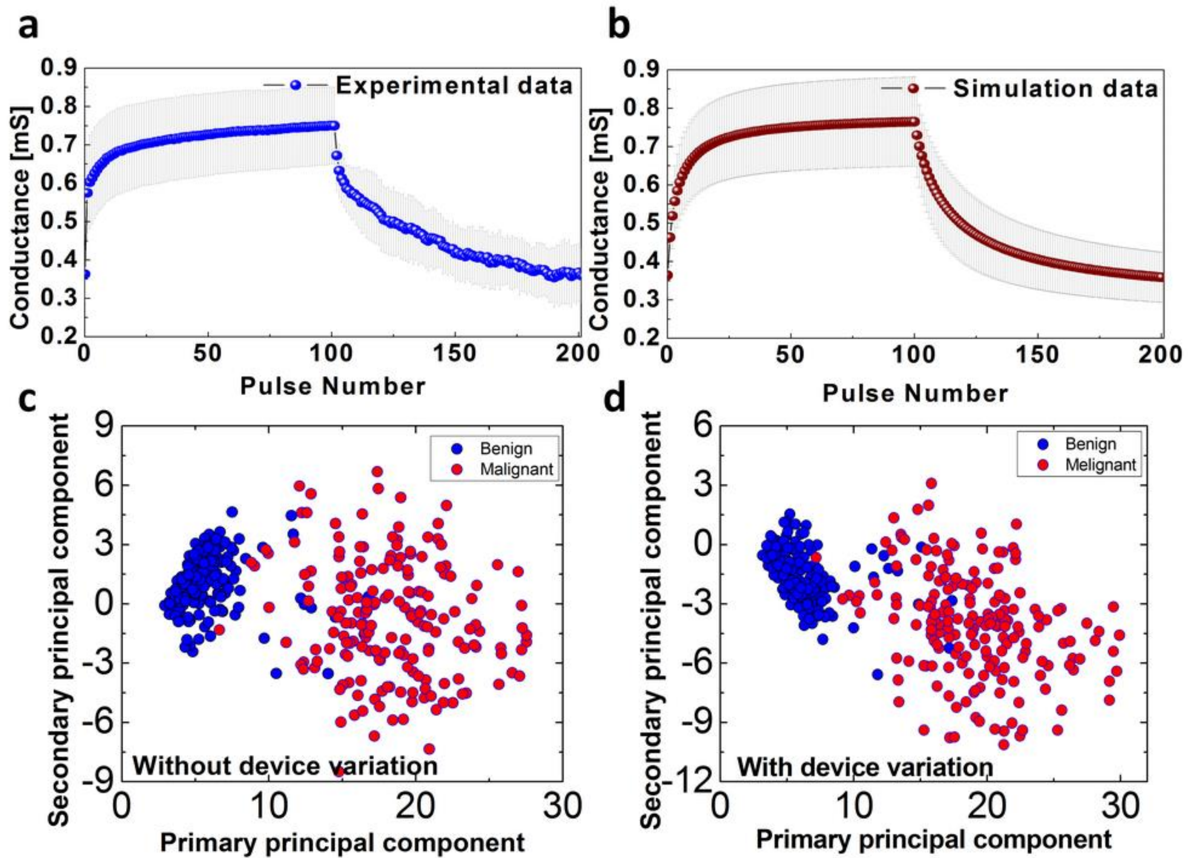


Figure 6.7. The effect of non-uniformity issue of the devices. (a) The measured data for the analog switching. The blue line and error bar represent the average and standard deviation, respectively. (b) Calculated analog behaviors adding the non-uniformity of the devices. (c) The result of the principal component analysis without device non-uniformity. (d) The result of the principal component analysis with device uniformity.

As shown in Fig. 6.9(a) and Fig. 6.9(b) (Appendix 6C), large device-device and cycle-cycle variations exist in the analog switching behaviors of RRAMs. The variations in the RRAM switching characteristics can be attributed to variations in device parameters such as the amount and distribution of oxygen vacancies in the conduction channel area, resistance variations of the TaO_x base region, stoichiometric non-uniformity and film thickness variations. Fig. 6.7(a) shows the conductance changes of 9 RRAM devices in the network during the application of 100 pulses of potentiation (-1 V) and 100 pulses of depression (1.15 V). The blue line represents the average value and the error bars represent the standard deviation of the measured conductance.

The relative standard deviation ranges from 10 % to 23 % for each point and are clearly substantial. To understand the effects of the device variations on the network performance, variations were introduced to the physical device parameters in Eqs. (1)-(2), and simulation results after incorporation of device variations are shown in Fig. 6.7(b), capturing the same average value and standard deviation as the measured data. Details of the measured data and modeling can be found in the Appendix 6C. The learning and PCA classification results of the RRAM network, with and without considering device variations, are shown in Fig. 6.7(c) and 6.7(d) for comparison. Significantly, even with substantial device-device and cycle-cycle variations (Fig. 7b), the network is still able to successfully learn the principal components and classify the data sets into the 2 categories (Fig. 6.7(d)). The training becomes slightly less optimal with the length of the weight vectors increased slightly to 1.05 and 1.06 for the primary and secondary principal components, respectively, compared to 1.0005 and 1.003 without considering device variations.

6.8 Analysis of Performance of the RRAM Network

Finally, to quantitatively analyze the performance of the RRAM network, logistic regression [25] was used to measure the effectiveness of the PCA analysis. The linear decision boundaries obtained from logistic regression are shown as dotted lines separating the two clustered sets of data in Fig. 6.7(c) and 6.7(d). Classification based on the PCA analysis and linear decision boundaries on the clustered data obtained from different approaches yielded essentially identical results (97.4% in Fig. 6.7(c) for the ideal case without considering device variations, and 97.6% in Fig. 6.7(d) for the case considering realistic device variations). This

result suggests that the RRAM network can be inherently tolerant to device variations due to the distributed network structure, and systems based on such networks can lead to reliable operations despite the nanoscale devices being intrinsically unreliable.

6.9 Conclusion

In conclusion, we show that RRAM networks can effectively implement unsupervised learning rules and be trained to learn principal components from data sets. The principal components learned during the training process can then be directly used to perform classification tasks using the same network. A realistic physical model was developed for the TaOx based RRAM and used in the analysis. Sanger's learning rule was utilized to implement online learning by adjusting the weights of each RRAM in the crossbar network. After learning the principal components, the RRAM network was successfully used to classify breast cancer data set as an example through first data clustering and then deriving a linear decision boundary. Significantly, successful learning and classification can still be obtained in the RRAM network even in the presence of substantial device variations, demonstrating the reliability of the network structure and the learning algorithm. The ability to achieve online learning and perform classification tasks reliably in the presence of unreliable devices suggest this approach can be extended to larger networks and other machine learning algorithms for more complex data-intensive tasks.

6.10 **Appendix**

6.10.1 Appendix A- Device Modeling

The growth rate of the state variable, *w*, is determined by ion hopping process over an energy barrier as shown in figure 6.8. It can be written as:

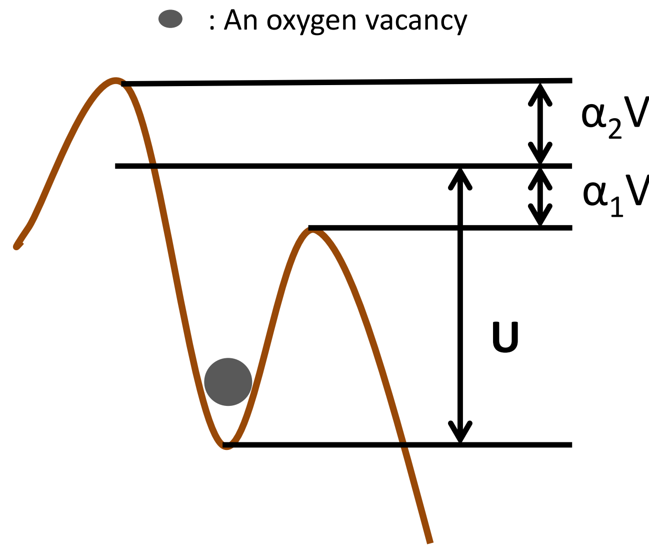


Figure 6.8. Energy barrier of ion hopping process.

$$\frac{dw}{dt} = f(w, V) \cdot d \cdot f\left[\exp\left(\frac{-q(U-\alpha_1 V)}{kT}\right) - \exp\left(\frac{-q(U+\alpha_2 V)}{kT}\right)\right] \quad (9)$$

where *d* is half of the average hopping distance of ions, *f* is the attempt frequency, *q* is the charge of an electron, *U* is the activation potential energy, *k* is the Boltzmann's constant, *T* is the temperature in Kelvin, α_1 and α_2 are barrier lowering coefficients [22], and *f(w,V)* is a window function to account for the non-linear response to the applied voltage[23]. The window function used in this paper is shown in equation (10):

$$f(w, V) = (w-1)^2 u(-V) + w^2 u(V) \quad (10)$$

Where *u()* is the Heaviside step function. By plugging equation (10) into equation (9), the rate equation of *w* can be re-written as:

$$\frac{dw}{dt} = (w-1)^2 k(e^{-\mu_1 V} - e^{\mu_2 V})u(-V) + w^2 k(e^{-\mu_1 V} - e^{\mu_2 V})u(V) \quad (11)$$

where $k = df\ exp(\frac{-qU}{kT})$, $\mu_1 = \exp(\frac{-q\alpha_1}{kT})$, and $\mu_2 = exp(\frac{-q\alpha_2}{kT})$. Eq. (11) is Eq. (1) in the main text.

The current through the device described by equation (2) consists of tunneling-dominated conduction and Schottky-dominated conduction. The tunneling current can be calculated by assuming MIM structure with very thin insulator so that the tunneling current is observed. Using the expressions for a square barrier [24-26], the current can be derived as:

$$I = A\frac{4q\pi m(kT)^2}{h^3}\exp(-b_1)\frac{1}{(c_1 kT)^2}\frac{\pi c_1 kT}{\sin(\pi c_1 kT)}(1-\exp(c_1 qV)) \quad (12)$$

Where

$$b_1 = \frac{2\alpha d\sqrt{q}}{3V}(\varphi_0^{\frac{3}{2}} - (\varphi_0 - V)^{\frac{3}{2}} \quad \text{if } V < \varphi_O \quad (13)$$

$$= \frac{2\alpha d\sqrt{q}}{3V}(\varphi_0^{\frac{3}{2}}) \quad \text{if } V > \varphi_O$$

$$c_1 = \frac{\alpha d}{V\sqrt{q}}(\varphi_0^{\frac{1}{2}} - (\varphi_0 - V)^{\frac{1}{2}}) \quad \text{if } V < \varphi_O \quad (14)$$

$$= \frac{\alpha d}{V\sqrt{q}}(\varphi_0^{\frac{1}{2}}) \quad \text{if } V > \varphi_O$$

A is the filament area, m is the effective electron mass, h is Plank's constant, φ_0 is the barrier height at zero bias, d is the tunneling distance, and $\alpha = \frac{4\pi\sqrt{2m}}{h}$.

At low bias, equation (13) and (14) can be simplified as:

$$b_1 = \frac{2\alpha d\sqrt{q}}{3V}\left(\varphi_0{}^{\frac{3}{2}} - (\varphi_0 - V)^{\frac{3}{2}}\right)$$

$$= \frac{2\alpha d\sqrt{q}}{3V}\varphi_O{}^{\frac{3}{2}}\left(1 - \left(1 - \frac{V}{\varphi_0}\right)^{\frac{3}{2}}\right)$$

$$= \frac{2\alpha d\sqrt{q}}{3V}\varphi_O{}^{\frac{3}{2}}(1 - \left(1 - \frac{3V}{2\varphi_0}\right))$$

$$= \frac{2\alpha d\sqrt{q}}{3V}\varphi_O{}^{\frac{3}{2}}(\frac{3V}{2\varphi_0}) \tag{15}$$

$$c_1 = \frac{\alpha d}{2\sqrt{q\varphi_0}} \tag{16}$$

By plugging equation (15) and equation (16) into the equation (12),

$$\mathrm{I} = \mathrm{A}\frac{4q\pi m(kT)^2}{h^3}\exp(-b1)\frac{1}{(c_1kT)^2}\frac{\pi c_1 kT}{\sin(\pi c_1 kT)}(1 - \exp(c_1 qV))$$

$$= \mathrm{A}\frac{4q\pi m(kT)^2}{h^3}\frac{1}{(c_1kT)^2}\frac{\pi c_1 kT}{\sin(\pi c_1 kT)}e^{-\alpha d\sqrt{q\varphi_0}}e^{\frac{\alpha dV}{4}\sqrt{\frac{q}{\varphi_0}}}\left(1 - e^{\frac{-\alpha dV}{2}\sqrt{\frac{q}{\varphi_0}}}\right)$$

$$= \mathrm{A}\frac{4q\pi m(kT)^2}{h^3}\frac{1}{(c_1kT)^2}\frac{\pi c_1 kT}{\sin(\pi c_1 kT)}e^{-\alpha d\sqrt{q\varphi_0}}\sinh(\frac{\alpha dV}{4}\sqrt{\frac{q}{\varphi_0}})$$

$$= \mathrm{A}\frac{16kT\pi^2 mq\sqrt{q\varphi_0}}{\alpha d h^3\sin\left(\frac{\pi\alpha dkT}{2\sqrt{q\varphi_0}}\right)}e^{-\alpha d\sqrt{q\varphi_0}}\sinh(\frac{\alpha dV}{4}\sqrt{\frac{q}{\varphi_0}}) \tag{17}$$

For the Schottky junction current is explained by equation (18).

$$I = \frac{qADn}{L}\left(1 - exp\left(-\frac{qV}{\vartheta kT}\right)\right) \tag{18}$$

where *q* in the charge of an electron, *A* is the dimension of the device, *D* is the diffusion coefficient, *n* is the number of electrons, *L* is the diffusion length of electrons, *V* is the applied voltage, ϑ is an ideality factor, *k* is the Boltzmann's constant and *T* is the temperature in Kelvin.

6.10.2 Appendix B- Normalization of the weights

The Sanger's rule originates from Hebb's rule . For the simplicity, with the assumption that we have only one output, we can write the Hebb's rule as shown in the equation below.

$$w_i(n+1) = w_i + \eta y x_i \tag{19}$$

If we require the weights to be normalized to prevent infinite growing output of Hebb's rule, the update rule is modified as:

$$w_i(n+1) = \frac{w_i + \eta y x_i}{\sqrt[2]{\sum_{j=1}^{m} (w_j + \eta y x_j)^2}} \tag{20}$$

where m is the number of inputs. Because η, the update rate, is normally very small(<<1), through Taylor expansion and keeping only the leading term, Equation (20) becomes.

$$w_i(n+1) = w_i + \eta y(x_{i-} w_i y) \tag{21}$$

Eq. (21) is the equation for Sanger's rule (with only 1 output. The case for more than one outputs can be derived similarly). In other words, by implementing Sanger's rule (21), we are effectively implementing rule (20) (again for small update rates η which is satisfied in experiments) with normalized weights.

6.10.3 Appendix C- Details of the measured data and modelling

Figure 6.10 shows the analog conductance changes of the 18 memristor devices forming the network. Device-device variations are clearly observed which causes conductance discrepancy after potentiation/depression pulses. To verify the effect of device variations on the network

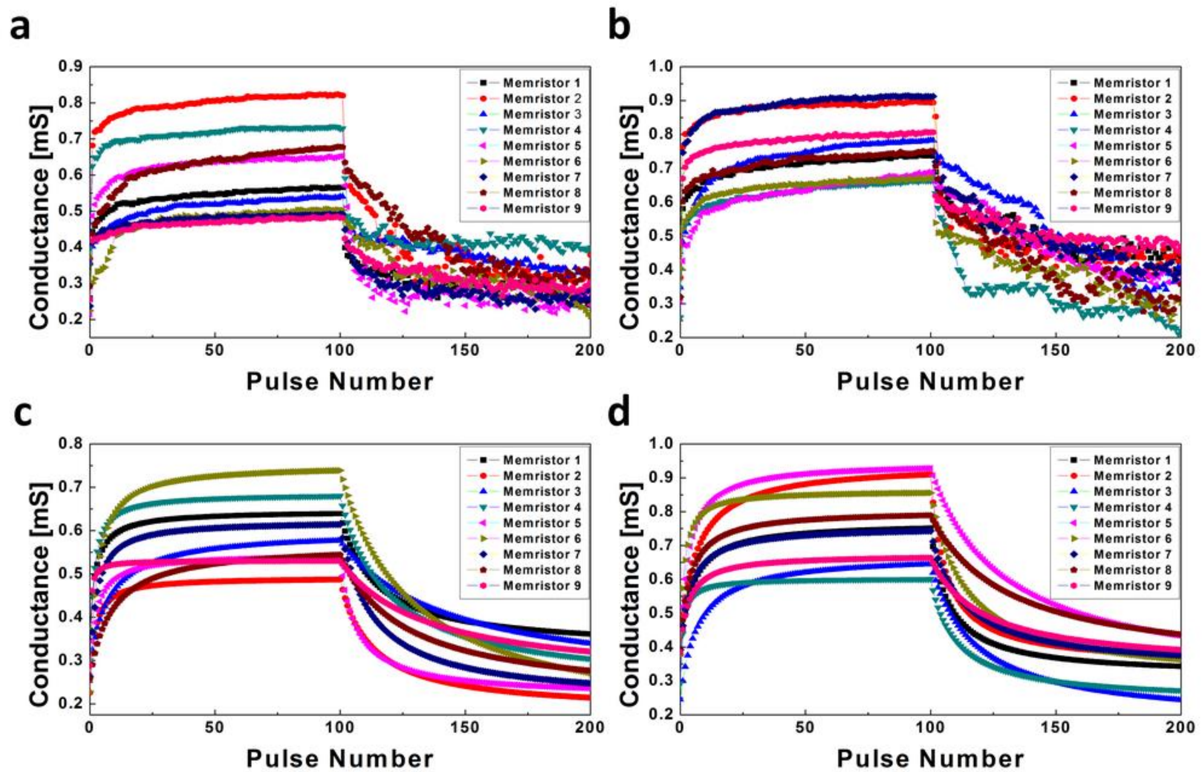


Figure 6.9. The details of conductance change measured at 0.2 V with 100 pulses of potentiation (-1 V) and 100 pulses of depression (1.15 V), consecutively for (a) measured conductance for 9 RRAM devices of the primary principal component (b) measured conductance for 9 RRAM devices of the second principal component (c) simulated conductance for 9 RRAM devices of the primary principal component (d) calculated conductance for 9 RRAM devices for the secondary principal component.

performance, the relevant device parameters were assumed to vary following Gaussian distributions (Table 6.1) and the exact value of a parameter for a given device was chosen randomly using a Monte Carlo method during the simulations. Figure 6.7(b) shows the average value and standard deviation calculated using this approach, which are consistent with the experimentally observed variations. The model with the random device variations was then applied to the network analysis and led to Figure 6.7(d). The parameters used in the simulation are shown in Table 6.1.

Variable	Primary Principal Component		Secondary Principal Component	
	Nominal value	σ/(Nominal value)	Nominal value	σ/(Nominal value)
κ	6×10^{-5}	3%	6×10^{-5}	3%
η_1	16	3%	16	3%
η_2	20.2	3%	20.2	3%
α	5×10^{-4}	15%	6×10^{-4}	15%
β	0.5	3%	0.5	3%
γ	2×10^{-3}	10%	2.55×10^{-3}	10%
δ	0.3	3%	0.3	3%

Table 6.1. Device parameters used in the simulation.

6.11 References

[1] Turel, Ö., Leem, J.H., Ma, X. & Likharev, K. K. Neuromorphic architectures for nanoelectronic circuits, Inter. J. Circuit Theory and Applications, 32, 277-302 (2004).

[2] Chua, L.O. & Yang, L. Cellular neural networks: theory. IEEE Trans. Circuits and Systems-I, 35, 1257–72 (1988).

[3] Jolliffe, I. T. Principal Component Analysis. (Springer, New York, 2002).

[4] Alibart, F., Zamanidoost, E. & Strukov, D. B. Pattern classification by memristive crossbar circuits using ex situ and in situ training. Nat. Commun. 4, 2072 (2013).

[5] Sheridan, P., Ma, W. & Lu, W. Pattern recognition with memristor networks. IEEE Inter. Symp. Circuits and Systems (ISCAS), 1078-81 (2014).

[6] Lee, M.-J. et al. A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O5-x/TaO2-x bilayer structures. Nat. Mater., 10, 625-30 (2011).

[7] Yang, J.-J. et al. High switching endurance in TaOx memristive devices. Appl. Phys. Lett., 97, 232102 (2010).

[8] Torrezan, A. C., Strachan, J. P., Medeiros-Ribeiro, G. & Williams, R. S. Sub-nanosecond switching of a tantalum oxide memristor. Nanotechnology, 22, 485203 (2011).

[9] Govoreanu, B. et al. 10 x 10nm2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation. Electron Devices Meeting (IEDM), 2011 IEEE Inter., 31–36 (2011).

[10] Yang, Y., Sheridan, P. & Lu, W. Complementary resistive switching in tantalum oxide-based resistive memory devices. Appl. Phys. Lett., 100, 203112 (2012).

[11] Strukov, D. B., Snider, G. S., Stewart, D. R. & Williams, R. S. The missing memristor found. Nature 453, 80–84 (2008).

[12] Yu, S., Guan, X. & Wong, H.-S. P. Conduction mechanism of TiN/HfOx/Pt resistive switching memory: A trap-assisted-tunneling model. Appl. Phys. Lett. 99, 063507 (2011).

[13] Chang, T., Jo, S., Kim, K.H., Sheridan, P., Gaba, S. & Lu, W. Synaptic behaviors and modeling of a metal oxide memristive device. Appl. Phys. A, 102, 857-63 (2011).

[14] Kim, S., Choi, S. & Lu, W. Comprehensive physical model of dynamic resistive switching in an oxide memristor. ACS Nano, 8, 2369-76 (2014).

[15] Mickel, P. R., Lohn, A. J., James, C. D. & Marinella, M. J. Isothermal switching and detailed filament evolution in memristive systems. Adv. Mater. 26, 4486–90 (2014).

[16] Choi, S., Lee, J., Kim, S. & Lu, W. D. Retention failure analysis of metal-oxide based resistive memory. Appl. Phys. Lett. 105, 113510 (2014).

[17] Kim, S., Choi, S., Lee, J. & Lu, W. D. Tuning resistive switching characteristics of tantalum oxide memristors through Si doping. ACS Nano, 8, 10262-69 (2014).

[18] Bache, K. & Lichman, M. Breast cancer Wisconsin (diagnostic) data set - UCI Machine Learning Repository (2013) Available at: http://archive.ics.uci.edu/ml. (Accessed: 6th July 2014)

[19] Wolberg, W. H. & Mangasarian, O. L. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc. Natl. Acad. Sci, 87, 9193–96 (1990).

[20] Oja, E. Simplified neuron model as a principal component analyzer. J. Mathmatical, 15, 267-73 (1982).

[21] Sanger, T. D. Optimal unsupervised learning in a single-layer linear feedforward neural network. Neural Networks, 2, 459-73 (1989).

[22] Bishop, C. M. Pattern recognition and machine learning [205] (Springer, New York, 2006).

[23] Ielmini, D. Modeling the Universal Set/Reset Characteristics of Bipolar RRAM by Field- and Temperature- Driven Filament Growth. *IEEE Trans. Elec. Dev.* 58**,** 4309–4317 (2011).

[24] Chang, T., Jo, S., Kim, K.H., Sheridan, P., Gaba, S. & Lu, W. Synaptic behaviors and modeling of a metal oxide memristive device. *Appl. Phys. A*, 102, 857-63 (2011).

[25] Kim, S., Choi, S. & Lu, W. Comprehensive physical model of dynamic resistive switching in an oxide memristor. *ACS Nano* 8, 2369–76 (2014).

[26] Stratton, R. Volt-current characteristics for tunneling through insulating films. *J. Phys. Chem. Solids* **23**, 1177-90 (1962)

[27] Simmons, J. G. Generalized Formula for the Electric Tunnel Effect between Similar Electrodes Separated by a Thin Insulating Film. *J. Appl. Phys.* **34**, 1793-1803 (1963)

Chapter 7.

Experimental Demonstration of Unsupervised Learning Using RRAM networks

7.1 **Introduction**

In chapter 6, we showed through simulation that RRAM networks can effectively implement unsupervised learning rules and be trained to learn principal components from data sets by simulation. As discussed, PCA is an important technique used in machine learning for preprocessing a data set or dimension reduction [1]. In this chapter, we study the experimental demonstration of unsupervised learning using RRAM crossbar arrays. The breast cancer data having malignant cell and benign cell is again used as an example in this experimental study.

7.2 **Device Fabrication and Analog switching behavior**

As discussed in chapter 5, our recent studies on RRAM devices having a tantalum oxide layer doped with silicon atoms show high On/Off ratio and controllable analog switching behavior [2]. Two sets of 16 by 1 arrays were fabricated and the device size is 2 µm × 2 µm as shown in Figure 7.1(a). The inset of figure 1b shows the schematic of the Pd/TaO_x/Ta_2O_5/Pd stack. As shown in figure 7.1(b), analog behavior is measured by a read (0.2 V) pulse during 100 potentiation pulses (-1 V, 10 µs) and 100 depression pulses (1.15 V, 10 µs). The plotted current is the average current from 9 devices.

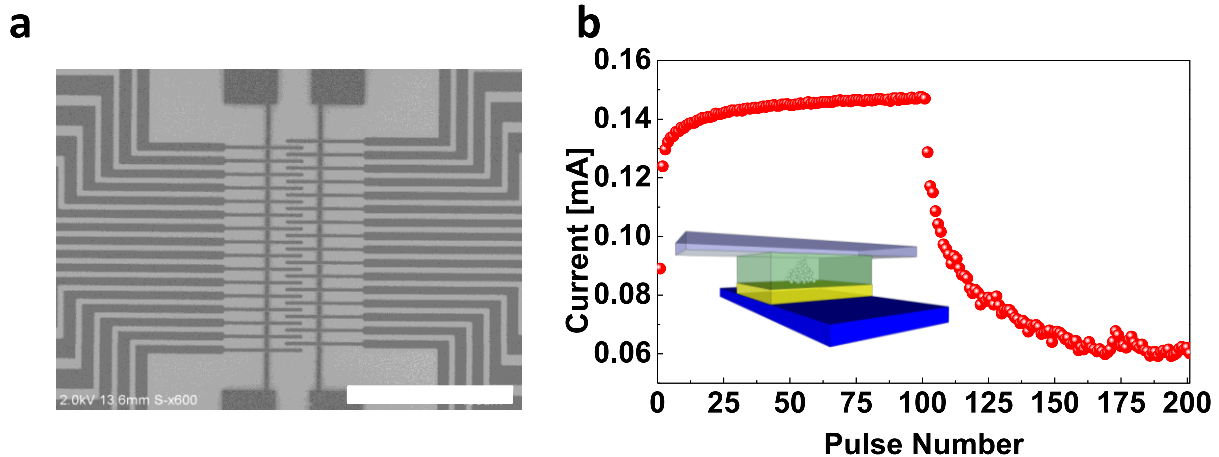


Figure 7.1 Device fabrication and analog switching behavior. (a) SEM images of the fabricated two sets of 16 by 1 RRAM devices. Scale bar: 100 µm. (b) DC I-V characteristics of a typical RRAM device showing the bipolar switching with 100 pulses of potentiation (-1 V, 10µs) and depression (1.15 V, 10µs), consecutively. Inset: schematic image of a RRAM device having oxygen vacancy filament. This is not to scale.

7.3 Peripheral Circuitry

To experimentally demonstrate the PCA network, a printed circuit board was designed by our collaborators (Prof. Zhengya Zhang's group) at University of Michigan, Ann Arbor. The board is shown in Figure 7.2(a). The Digital to Analog Converters (DACs) (Analog Devices, model AD7305) provide bias to the top electrodes (TEs) and bottom electrodes (BEs). The switches (Analog Devices, model ADG 738) are utilized to select a cell so the cell can be programmed or be read out. The Analog to Digital Converter (ADC) (Linear Technology, model LTC 1412) and Op amp (Texas Instruments, model OPA 657) are designed to read current or calculate charge during read operation. The left part of the board is connected to a microcontroller with FPGA chip (Xilix, Spartan 6) for the communication between a computer

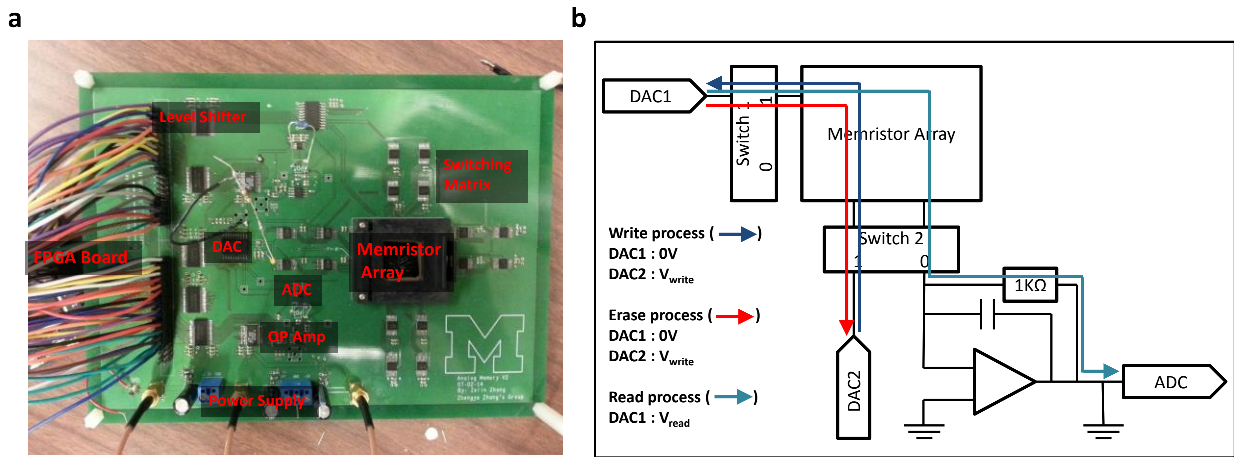


Figure 7.2 Peripheral circuitry (a) the photo image of the board with label parts. (b) schematic of the procedure of the board operation.

and the board. A wire bonded RRAM array shown in Figure 7.1(a) is attached to the board through a socket.

The detailed procedure of the board operation is shown in figure 2b. For the write process, a positive write voltage is applied to DAC2 while DAC1 is grounded so BE has positive voltage and TE is grounded. Under this condition, a selected RRAM device is subjected to an effectively negative write voltage at the TE and 0 V at the BE. The oxygen vacancies migrate in the direction of TE, so the device becomes more conductive. For the erase process, a positive erase voltage is applied to DAC1 with 0V applied to DAC2, so the oxygen vacancies in the filament of the selected device move back to the BE resulting in decrease of conductivity of the RRAM device. For the write and erase process explained above, the switch matrix 1 and switch matrix 2 have value 1 for selected devices to apply voltages between DAC1 and DAC 2. For read process, DAC1 applies the read voltage and the current is read using the op amp, a sensing

resistor and the ADC. The switch matrix 2 output value 0 while the switch matrix1 output value 1 for the selected devices to get current from DAC1 to ADC.

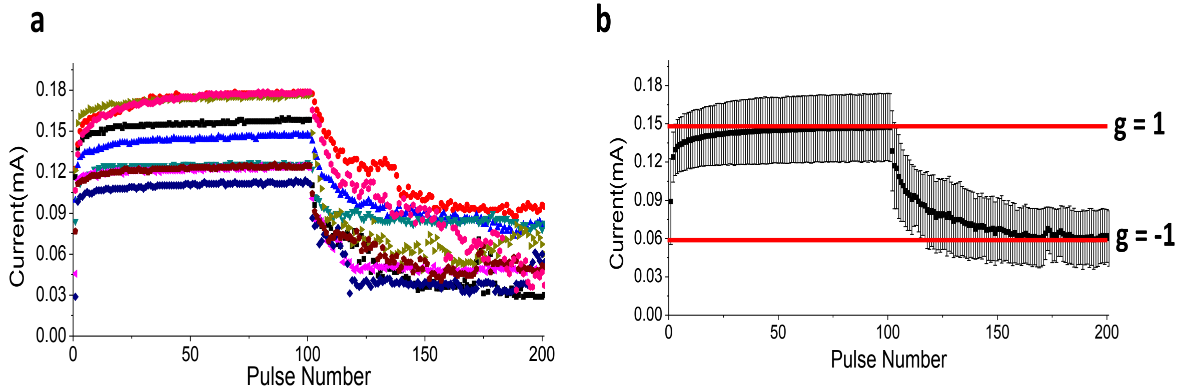


Figure 7.3 (a) Experimental measurements collected by the board for 9 RRAM devices in the same column (corresponding to the second principle component), showing the analog conductance change and device-device variations. The conductance was measured with 0.2 V, 1 ms pulses, and the devices were subject to 100 pulses of potentiation (-1 V, 10 µs) and 100 pulses of depression (1.15 V, 10 µs). (b) The solid line and the error bars represent the average and standard deviation.

7.4 Learning in the RRAM array

As discussed in chapter 6, even with the presence of substantial device-device and cycle-cycle variations, the network is still able to successfully learn the principal components and cluster the data sets. Figure 7.3(a) shows responses from each RRAM devices measured by 0.2V read pulse when 100 potentiation and 100 depression pulses applied. To obtain the weights g_{ij} for each device, Eq. (1) was used along with the measured current I through the device at applied voltage V. Same parameters (γ, δ, α, and β) were used for all devices, and those parameters were obtained by fitting the memristor equations with the averaged measured current.

$$I = \frac{1+g}{2}\gamma \sinh(\delta \times V) + \frac{1-g}{2}\alpha\left(1 - e^{-\beta \times V}\right) \quad (1)$$

To initialize the PCA network, initial weights of each RRAM devices were calculated by applying read voltage pulse. From the current equation given by equation (1), weight, g, is extracted.

During training, voltage pulses representing input *x*, are applied to the inputs (rows). Charge is calculated from the measured current at each output (column), and the output function *y* at each column is calculated through equation (2)

$$y_j = \frac{2Q_j}{A-B} - \sum_i [\frac{A+B}{A-B} x_i] \quad (2)$$

As shown in Ch. 6, $y_j = \sum_{i=1}^{n} g_{ij} x_i$ following the definition used in (2).

To learn the principal components from the training set, Sanger's rule is applied to calculate the changes of weights following equation (3) [3].

$$\Delta g_{ij} = \eta y_j (x_i - \sum_{k=1}^{j} g_{ik} y_k) \quad (3)$$

As discussed in Ch. 6, to implement Sanger's rule, the programming pulse width, $|\Delta t|$, is calculated based on the current weight and Δg as shown in equation (4).

$$\Delta t_{ij} = f(g_{ij}, \Delta g_{ij}) = \frac{2}{k\left(e^{-\mu_1 V_{potentiation}} - e^{\mu_2 V_{potentiation}}\right)} \left(\frac{-1}{g_{ij,after} - 1} + \frac{1}{g_{ij,before} - 1}\right) u(\Delta g)$$

$$+ \frac{2}{k\left(e^{-\mu_1 V_{depression}} - e^{\mu_2 V_{depression}}\right)} \left(\frac{-1}{g_{ij,after} + 1} + \frac{1}{g_{ij,before} + 1}\right) u(-\Delta g) \quad (4)$$

The details of the calculation of the parameters above are explained in chapter 6.

Since the practical minimum pulse width with the board is around 700ns, in our experiment, 1μs pulse width is used if the calculated pulse width is between 100ns and 1μs, and the pulse is

ignored if the calculated pulse width is smaller than 100ns. Once the pulse width is calculated, the actual write pulse or erase pulse is applied to the device. A write pulse with pulse width, Δt, is applied when Δg is positive while an erase pulse with pulse width, Δt, is applied when Δg is negative. After applying all training pulses to the RRAM array, the updated weight is calculated by applying a read voltage. The procedure explained above is then repeated until all training data are used in one cycle of training. To apply another cycle of training, the sequence of the training

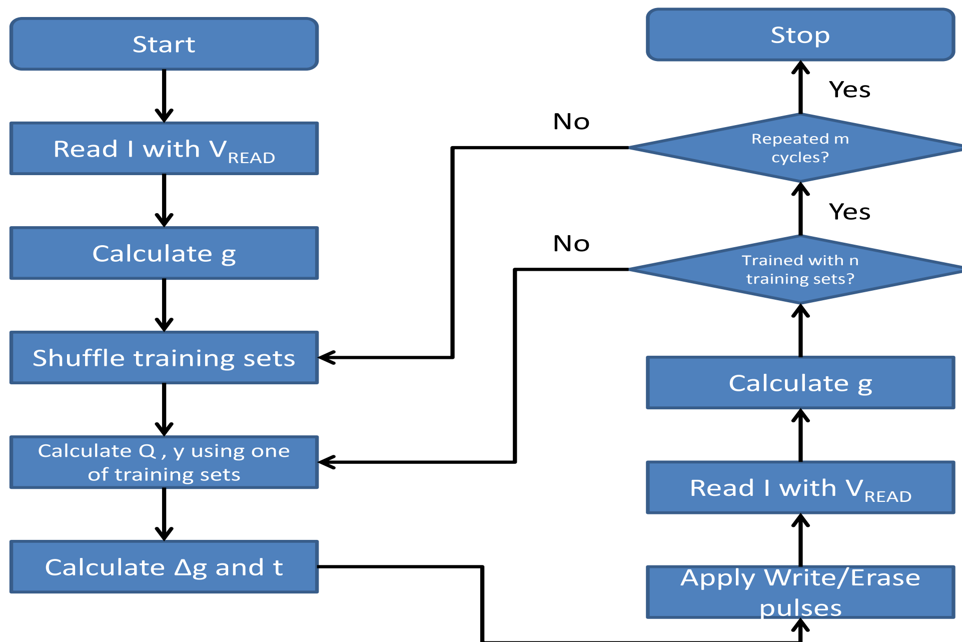


Figure 7.4 Flowchart showing the overall operation procedure.

data is randomized to prevent the RRAM array from learning the (artificial) pattern of sequence of the training data. The flowchart in figure 7.4 is the summary of the operating procedure explained above.

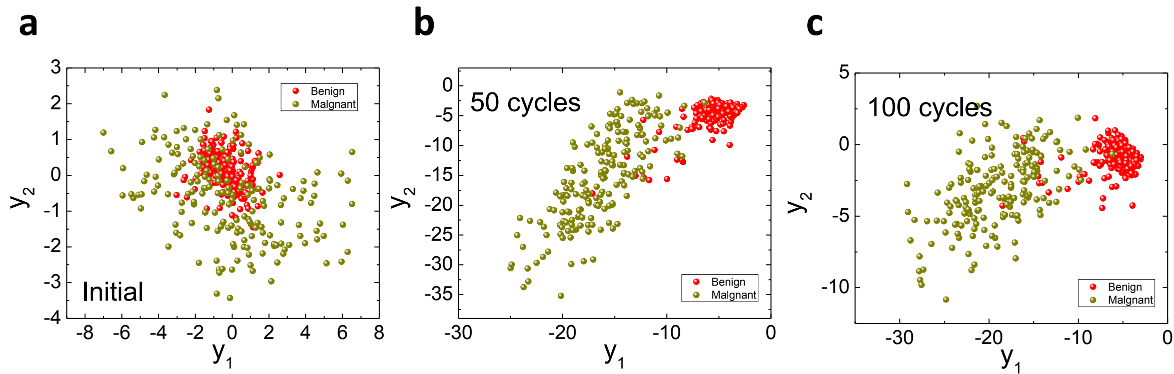


Figure 7.5 Results of principal component analysis. The data are plotted on y_1 and y_2 axis. (a) Initial results of an untrained RRAM network. (b) Results of a partially trained RRAM network. (c) Results of a fully trained RRAM network.

7.5 Data Clustering before and after Applying Unsupervised Learning Rule

To test the PCA network, a standard breast cancer data set from University of Wisconsin Hospitals, Madison was adopted as the input signal data, as discussed in chapter 6. Each training cycle consists of one hundred randomly sequenced data points (50 points from benign class, 50 points from malignant class). After training, the ability of the network to cluster the data was tested by analyzed 583 data points from the same data base but not included in the training data. Figure 7.5(a) shows the output from the network for the 583 data points before learning, with y_1 as the horizontal axis and y_2 as the vertical axis. Red dots and dark yellow dots show benign and malignant cells (ground truth), respectively. Before training, the two groups, red dots and dark yellow dots are mixed with each other. In other words, the network was not able to cluster the data without training. Figure 7.5(b) shows the results after 50 cycles of training. As explained in chapter 6, for the primary principal component (y_1), the weights changes rapidly and become stabilized while the weight changes for the secondary principal component (y_2) needs more time to be stabilized. The Euclidean norm of the weights for the primary principal component is 1.06

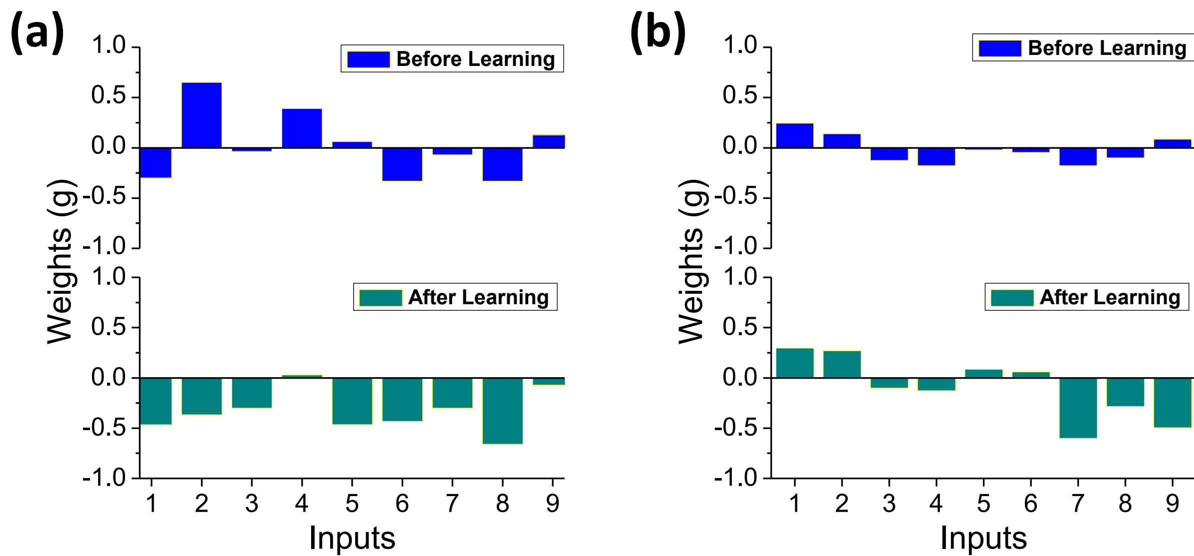


Figure 7.6 Weights constituting (a) the primary principal component and (b) the secondary principal component before (upper graph) and after (lower graph) the learning process.

which is almost converged to 1 (as expected from implementing Sanger's rule, Ch 6) and that of the weights for the secondary principal component is 1.56 in figure 7.5(b). This implies that the weights for the primary principal component are stabilized while the weights for the secondary principal component still need to trained further. Figure 7.5(c) shows the results after 100 cycles of training. The Euclidean norm of the weights for both primary principal component and secondary principal component are closed to 1, indicating the network has completed learning.

As shown in figure 7.6, the weight distributions changed for both primary principal component and secondary principal component during the learning step. Interestingly, the weights for the primary principal component in this experimental study after learning were found to be all negative, while the weights found in our simulation in chapter 6 were all positive. The primary principal component corresponds to the direction of the largest variance in the data set, and the experimentally obtained primary principle component in Fig. 7.6 is roughly the mirror

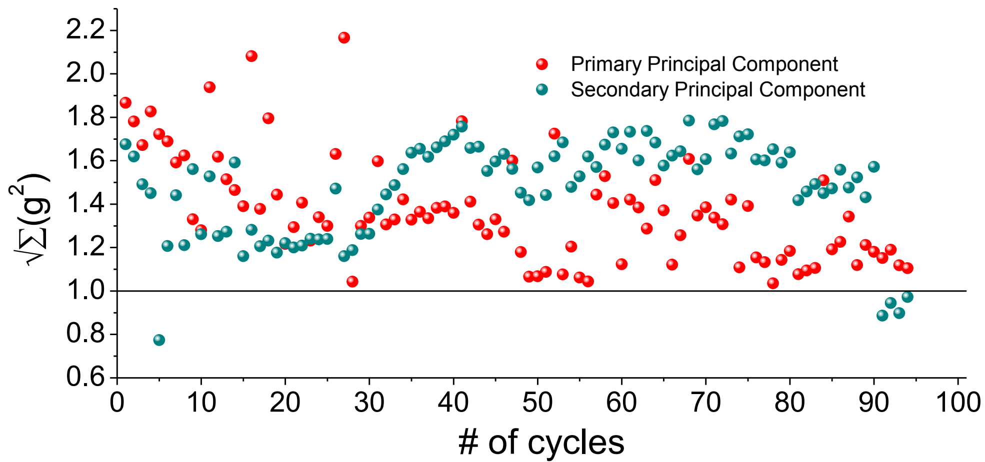


Figure 7.7 Evolution of the Euclidean norm of weights during learning. Red dots represents the norm value of the weights for the primary principal component and dark cyan dots shows the norm value of the weights for the secondary principal component.

image of that obtained through simulation shown in Fig. 6.4, so they in fact correspond to (roughly) the same direction and leading to (roughly) the same primary principle component. However, those two results are not a perfect match of each other, since Sanger's rule only leads to the exact principle components after infinite training in ideal cases, and the actual vectors obtained in both experimental and simulation studies are only approximating the exact solutions. However, the approximation is still sufficiently close enough to the exact solutions and excellent clustering and classification results can still be obtained.

Figure 7.7 shows the evolution of the Euclidean norm of the weights during learning. The red dots and the dark cyan dots represent the norm value for primary principal component and secondary principal component, respectively. It can be seen that the norm value of the primary principal component converges to 1 faster than that of the secondary principal component. This is because the convergence of the secondary principal component only happens after the primary principal component has converged. It is noted that the norm values are not

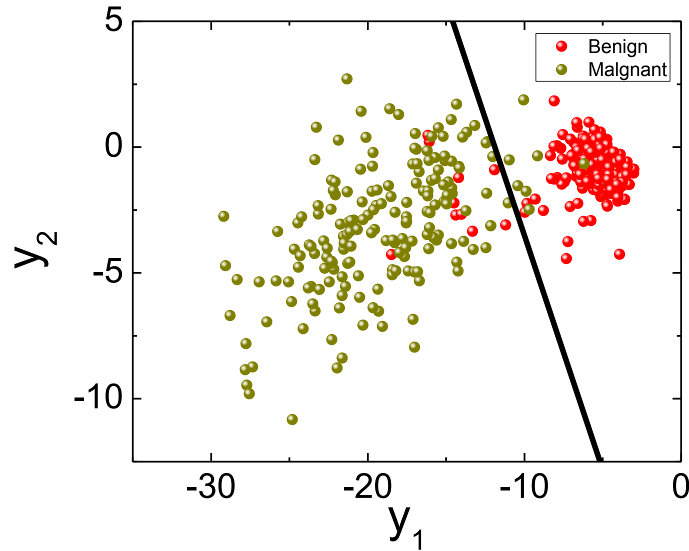


Figure 7.8 Classification based on linear decision boundary (black line) on the clustered data.

stabilized and some points are out of the tendency. For example, devices can be stuck at the highest current level or lowest current level of the device which makes the norm value high and out of the trend. This problem can be resolved by applying longer write or erase pulse for the devices stuck at low current level or high current level, respectively.

To analyze the clustered data to measure the accuracy of the RRAM network, a linear decision boundary was developed by logistic regression [4]. The decision boundary obtained from logistic regression is shown as black line in figure 7.8. The boundary line separates the two sets of clustered data accurately. Only 17 data points among 583 data points are misclassified corresponding to 97% accuracy. This result suggests that data clustering through an RRAM-based network employing unsupervised learning can be used for effective data classification.

7.6 Conclusion

In conclusion, we show that RRAM networks can implement PCA, one of the most used and representative unsupervised learning rule, and successfully cluster data in a real-world environment. The experiments were carried out in TaO_x-based RRAM arrays and a customized PCB board with FPGA and a microcontroller. Online learning was successfully demonstrated by adjusting the weights of each RRAM device in the crossbar network through unsupervised learning even with abnormal unpredictable behavior of the devices. We believe this experiment contributes to bridging the software based machine learning algorithm to hardware framework.

7.7 **References**

[1] Jolliffe, I. T. Principal Component Analysis. (Springer, New York, 2002).

[2] Kim, S., Choi, S., Lee, J. & Lu, W. D. Tuning resistive switching characteristics of tantalum oxide memristors through Si doping. ACS Nano, 8, 10262-69 (2014).

[3] Sanger, T. D. Optimal unsupervised learning in a single-layer linear feedforward neural network. Neural Networks, 2, 459-73 (1989).

[4] Bishop, C. M. Pattern recognition and machine learning [205] (Springer, New York, 2006).

Chapter 8.
Summary and Future Work

8.1 Discussion

In chapter 2, we studied modeling of the dynamic resistive switching processes in RRAM. We developed a compreshensive and accurate physical model that quantitatively explains the dynamic memristive switching process by solving the local electric field, temperature and V_O concentration self-consistently. This model confirms that the conductive filament is formed and ruptured inside the switching layer. The set process involves field-driven filament formation followed by filament expansion, while reset process is dominated by thermal-driven filament rupture followed by gap widening. A quantitative and accurate dynamic switching model that fully accounts for the resistive switching behaviors in RRAM in a unified framework provides a physical picture of the resistive switching behavior and a basis for continued device optimizations.

In chapter 3 and chapter 4, we discussed several experiments that revealed the resistive switching mechanism and distributions of oxygen vacancies inside the switching layer. In chapter 3, we carried out systematic investigations of the mechanism using detailed noise analysis and electron transport analysis. These systematic analyses verified the conduction channel formation is associated with the distribution of oxygen vacancies. Interestingly, as the device was switched from HRS to LRS the conduction channel area was reduced although the

local oxygen vacancy concentration is increased. In chapter 4, we studied retention failure mechanism of RRAM devices at high temperature. The activation energy for oxygen vacancy diffusion was calculated and analyzed by both analytical modeling and detailed numerical multi-physics simulation, which verifies filament-based conduction path in LRS.

In chapter 5, we studied how to improve the resistive switching behavior in RRAM devices. By systematical tuning at the atomic level through doping, the resistive switching characteristics were improved. Si doping expedited oxygen vacancy drift and the device showed larger dynamic analog switching ranges with high endurance. These findings will provide valuable information on the application of neuromorphic computing system.

In chapter 6 and chapter 7, we investigated the application of the RRAM in neuromorphic computing, by showing data clustering based on unsupervised learning. In chapter 6, we investigated PCA network using RRAM crossbar arrays through simulations with realistic device models and also accounting for expected device variations. The trained RRAM network was able to cluster and separate input data set into 2 categories. Even with relatively large device variability, the network with RRAM devices categorized the input signals with high accuracy. In chapter 7, the concept of RRAM-array based PCA network was demonstrated experimentally. Detailed experimental procedures have been developed and successful data clustering and classification were achieved through the trained RRAM network and peripheral circuitry at the board level.

8.2 **Future work**

8.2.1 RRAM Crossbar Array for Preprocessing of Neural Signal

In chapter 6, we demonstrated the feasibility of performing component analysis based on RRAM devices. In principle, the RRAM-based hardware can be directly embedded in biological systems due to its simplicity, high endurance [1,2] and low power consumption [3]. For the next step, coupling of the artificial neural network system with biological systems will be demonstrated. From a set of bio neural signal collected by electroencephalogram (EEG), electrocardiogram (ECG) or electromyogram (EMG), the pulse data can be obtained as shown in Fig. 8.1. Fig. 8.1(a) shows an example of the measured data from a channel in motor cortex of a

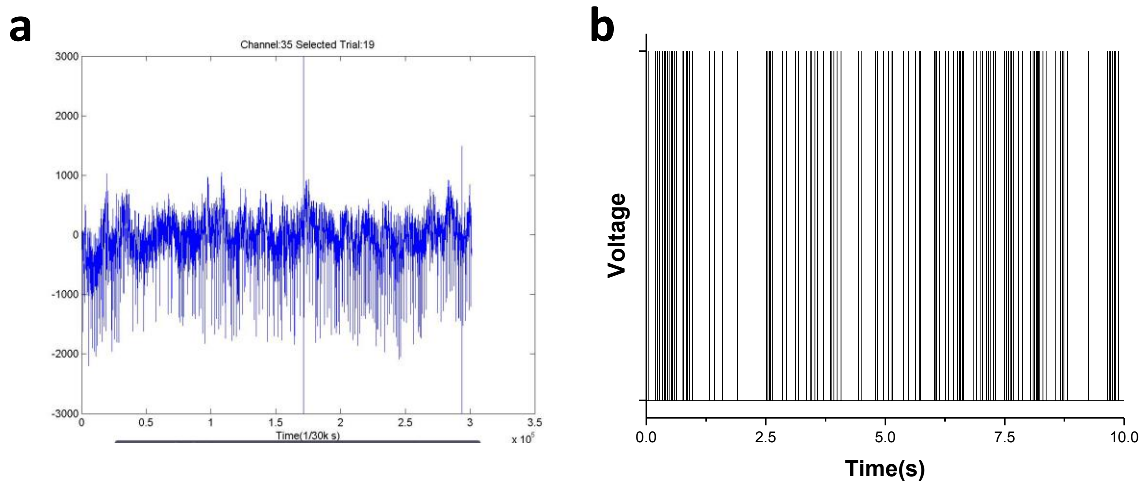


Figure 8.1. (a) Measured neural signals in motor cortex from a Monkey. Data obtained from Cortical Neural Prosthetics Lab from Biomedical Engineering in University of Michigan. (b) Extracted spikes with constant voltage from Fig. 1(a).

Monkey. The neural signal data consist of pulses collected from 96 channels. By stimulating individual fingertips, the fire rate at each neuron (measured by each channel) would be modulated. From the recorded data, we can extract the (digitized) spike patterns as shown in Fig.

8.1(b) by applying the equation, $V_{TH} = -4.5 \times V_{RMS}$ [4]. When voltage is higher than V_{TH}, it is considered as spike. The calculated spike firing pattern serves as the input signal of the RRAM arrays to perform PCA network using Sanger's rule. The unsupervised learning discussed in chapter 6 and chapter 7 may enable a neural signal processing system embedded the body that can efficiently and quickly analyze biological neural signals for potential diagnosis and prosthesis applications.

8.2.2 Device Optimization – Analog Switching

As discussed in Chapters 6 and 7, analog switching behavior of RRAM devices can be utilized for neuromorphic computing systems where memory and logic are implemented at the same physical locations. However, reliable analog switching behaviors of RRAM devices still need further development. For example, tungsten oxide based RRAM devices [5] show reliable analog switching, but they normally offer short retention times that limit their range of applications. The Si doped tantalum oxide based RRAM device discussed in chapter 5 shows good retention and long endurance, but they still suffer issues such as device to device and cycle to cycle uniformity variations and relatively small window for the analog switching. Optimizing the analog switching characteristics is a relatively unexplored areas and lots of work still need to be performed to identify and optimize the switching material and engineer the device stack. Continued optimization of analog RRAM devices will not only leads to better memories but can also result in reliable analog computing systems or neuromorphic computing systems based on RRAM devices.

8.3 References

[1] Lee, M. -J.; Lee, C. B.; Lee, D.; Lee, S. R.; Chang, M.; Hur, J. H.; Kim, Y. –B.; Kim, C. –J.; Seo, D. H.; Seo, S.; Chung, U. –I.; Yoo, I. –K.; Kim, K. A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O5-x/TaO2-x bilayer structures. Nat. Mater. 2011, 10, 625-630.

[2] Kim, S.H.; Choi, S.H.; Lee, J.H.; Lu, W. D. Tuning Resistive Switching Characteristics of Tantalum Oxide Memristors through Si Doping. ACS Nano. 2014, 8(10), pp 10262-10269.

[3] B. V. V Zhirnov, R. K. Cavin, L. F. Ieee, S. Menzel, E. Linn, S. Schmelzer, D. Bra, C. Schindler, and R. Waser, “Memory Devices : Energy – Space – Time Tradeoffs,” Proc. IEEE, 2010, 98(12), pp. 2185–2200.

[4] Chestek, C. A; Gilja, V.; Nuyujukian, P.; Foster, J. D.; Fan, J. M.; Kaufman, M. T.; Churchland, M. M.; Rivera-Alvidrez, Z.; Cunningham, J. P.; Ryu, S. I.; et al. Long-Term Stability of Neural Prosthetic Control Signals from Silicon Cortical Arrays in Rhesus Macaque Motor Cortex. J. Neural Eng. 2011, 8, 045005.

[5] Chang, T.; Jo, S.-H.; Kim, K.-H.; Sheridan, P.; Gaba, S.; Lu, W. Synaptic Behaviors and Modeling of a Metal Oxide Memristive Device. Appl. Phys. A 2011, 102, 857–863.