Controlling Ionic Transport in RRAM for Memory and Neuromorphic Computing Applications

by

Jihang Lee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Materials Science and Engineering)
in the University of Michigan
2018

Doctoral Committee:

Professor Wei D. Lu, Co-Chair
Associate Professor Emmanouil Kioupakis, Co-Chair
Associate Professor Pierre Ferdinand P. Poudeu
Associate Professor Zhaohui Zhong

Jihang Lee

jihang@umich.edu

ORCID ID: 0000-0001-8214-5905

Acknowledgements

First, I would like to express my sincere gratitude to my advisors Prof. Wei D. Lu and Prof. Emmanouil Kioupakis for their continuous support on my Ph.D. study and related research. They have given me insightful and thoughtful advices with their broad vision/knowledge and great patience throughout my PhD study. Both advisors have always been willing to devote their time to interacting and helping me. I have truly enjoyed discussions and conversations with them, which enabled me to overcome challenging issues in my research. In addition to their tremendous help in research, they have set a best model with their faithfulness, kindness, and sincerity. I am fortunate to have not one or two advisors as a good mentor in my scientific research and my life.

I would also like to thank you my committee members, Prof. Pierre Ferdinand P. Poudeu and Prof. Zhaohui Zhong, for their membership, and for sharing their knowledge and time with useful discussions.

My sincere thanks also go to my former and current group members: Dr. Yang, Dr. Moon, Dr. Kim, Dr. Gaba, Dr. Choi, Dr. Sheridan, Dr. Lin, Dr. Zhou, Dr. Du, Dr. Otuonye, Dr. Ma, Dr. Zhu, Dr. Zidan, Fuxi Cai and Yeonjoo Jeong, Jonghoon Shin, Seung Hwan Lee, John Moon, William Schell, Qiwen Wang, and Fan-Hsuan Meng in Prof. Lu's group and Dr. Bayerl, Dr. Shi, Dr. Waters, Dr. Jones, Dr. Olvera, Andrew McAllister, Logan Williams, Kelsey Mengle, Nocona Sanders, Sieun Chae, and Zihao Deng in Prof. Kioupakis'group for friendship, assistance and memorable moments.

I would like to express my deepest gratitude to my parents, family, and friends for their relentless support and encouragement during my course of study and also colleagues in Materials

Science and Engineering and Electrical Engineering and the Nanofabrication Facility (LNF) staff for their support.

Lastly, I acknowledge funding sources supported my graduate research including Kwanjeong Scholarship and Rackham Predoctoral Fellowship.

Table of Contents

List of Figures

Abstract

Resistive random-access memory, based on a simple two-terminal device structure, has attracted tremendous interest recently for applications ranging from non-volatile data storage to neuromorphic computing. Resistive switching (RS) effects in RRAM devices originate from internal, microscopic ionic migration and the associated electrochemical processes which modify the materials' chemical composition and subsequently their electrical and other physical properties. Therefore, controlling the internal ionic transport and redox reaction processes, ideally at the atomic scale, is necessary to optimize the device performance for practical applications with large-size arrays. In this thesis we present our efforts in understanding and controlling the ionic processes in RRAM devices.

This thesis presents a comprehensive study on the fundamental understanding on physical mechanism of the ionic processes and the optimization of materials and device structures to achieve desirable device performance based on theoretical calculations and experimental engineering.

First, we investigate the electronic structure of Ta_2O_5 polymorphs, a resistive switching material, and the formation and interaction of oxygen vacancies(V_{OS}) in amorphous Ta_2O_5, an important mobile defect responsible for the resistive switching process, using first-principles calculations. Based on the understanding of the fundamental properties of the switching material and the defect, we perform detailed theoretical and experimental analyses that reveal the dynamic

V_O charge transition processes, further helping the design and optimization of the oxide-based RRAM devices.

Next, we develop a novel structure including engineered nanoporous graphene to control the internal ionic transport and redox reaction processes at the atomic level, leading to improved device performance. We demonstrate that the RS characteristics can be systematically tuned by inserting a graphene layer with engineered nanopores at a V_O-exchange interface. The amount of V_Os injected in the switching layer and the size of the conducting filaments can be effectively controlled by the graphene layer working as an atomically-thin ion-blocking material in which ionic transports/reactions are allowed only through the engineered nanosized openings.

Lastly, better incremental switching characteristics with improved linearity are obtained through optimization of the switching material density. These improvements allow us to build RRAM crossbar networks for data clustering analysis through unsupervised, online learning in both neuromorphic applications and arithmetic applications in which accurate vector-matrix multiplications are required. We expect the optimization approaches and the optimized devices can be used in other machine learning and arithmetic computing systems, and broaden the range of problems RRAM based network can solve.

Chapter 1 Introduction

1.1 On-Demand Reconfiguration of Nanomaterials

In conventional electronic devices such as transistors, the information is processed by controlling the mobile carriers, e.g., electrons, schematically shown in Figure 1.1. Ions such as dopants are necessary to build the device during fabrication, but are not actively modulated/programmed during device operation. Thus, no physical changes to the devices and materials are involved during device operation. On the contrary, physical reconfiguration of a material requires moving masses (atoms or ions) during the device operation. Recent advances in thin-film and fabrication technologies, combined with developments in characterization and

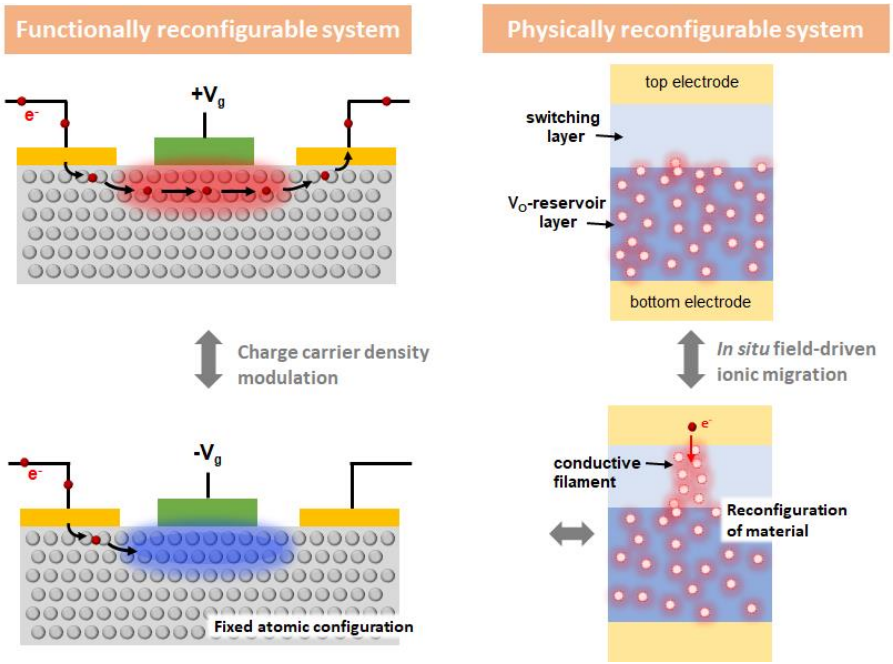

Figure 1.1: Schematic illustration showing how functionally (left) and physically (right) reconfigurable systems are differently operated. Reproduced with permission.[2] Copyright 2017, John Wiley and Sons.

measurement techniques, have now made it possible to control the migration of individual atoms (ions) in even conventional insulators and semiconductors.[1,2] These devices can be programmed at high speed, with other operation parameters also compatible with conventional electronic circuits, and thus enable the merge of ionics with electronics.

1.2 Resistive Random Access Memory (RRAM)

Ionic-based resistive switching devices that operate by controlled physical reconfiguration of the switching material, schematically shown in Figure 1.1, have been extensively studied with excellent performance metrics.[1,4] Specifically, these devices possess the simplicity of a resistor – with the whole device structure consisting of only two electrodes sandwiching a "switching" medium. Unlike a simple resistor, however, the device offers the capability to change its resistance on demand, and more importantly, the ability to store (memorize) the new resistance state as needed, thus the name memristor (memory + resistor)[5-7], or RRAM in memory applications. The resistance change and storage (termed resistive switching, RS) effects are achieved by the redistribution of ions (atoms) in the switching layer (Figure 1.1), which in turn modulate the local

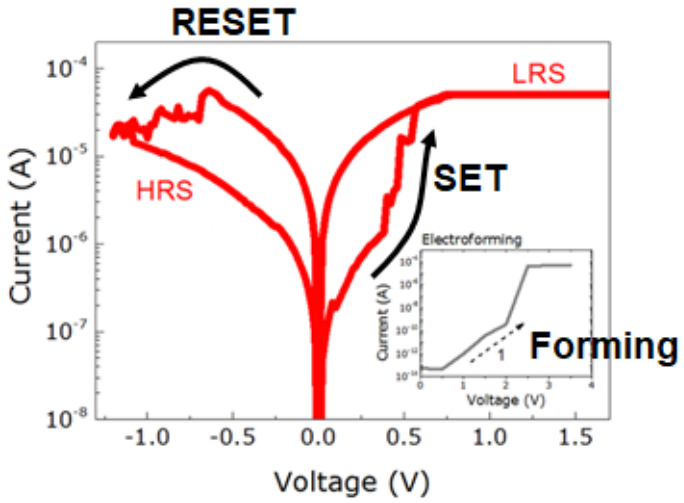

Figure 1.2: Resistive switching characteristics in a RRAM device. Reproduced with permission.[3] Copyright 2015, American Chemical Society.

resistivity and the overall electrical resistance, as shown in Figure 1.2.[1,4] For example, a RRAM device can be programmed from high resistance state to low resistance state (SET process) by applying a voltage bias and then the device resistance can be switched back to the high resistance state (RESET process) by applying opposite polarity bias. Each state is maintained until another set or reset bias is applied, indicating non-volatile switching. In this case, the functionality, e.g. memory effect as reflected by the resistance changes, is realized through ionic processes but the information is sensed and processed through electrical signals, achieving efficient data storage and computing functions in an extremely compact structure. Therefore, understanding and controlling the internal ionic transport and redox reaction processes, ideally at the atomic scale, is necessary to optimize the device performance for practical applications with large-size arrays.

1.3 RRAM for High-Density Data Storage

A main feature of RRAM is the high storage-density the device offers. Specifically, RRAM's simple, two-terminal structure (Figure 1.3a) enables device integration in a crossbar form with high-density and high-connectivity, having one device formed at each crosspoint in the crossbar, as shown in Figure 1.3b. In the crossbar structure, each device can be randomly accessed and the cell size is minimum, i.e. $4F^2$ (F representing the smallest feature size), allowing highest possible 2D density. Additionally, RRAM fabrication typically involves only a few additional steps, using low-temperature processes and materials that are mostly compatible with complementary-metal-oxide-semiconductor (CMOS) processes. This makes it possible to directly integrate RRAM arrays on top of CMOS circuitry in the same chip (Figure 1.3c), as well as 3D stacked memory structures with even higher density (Figure 1.4). Combined with the sub-10nm

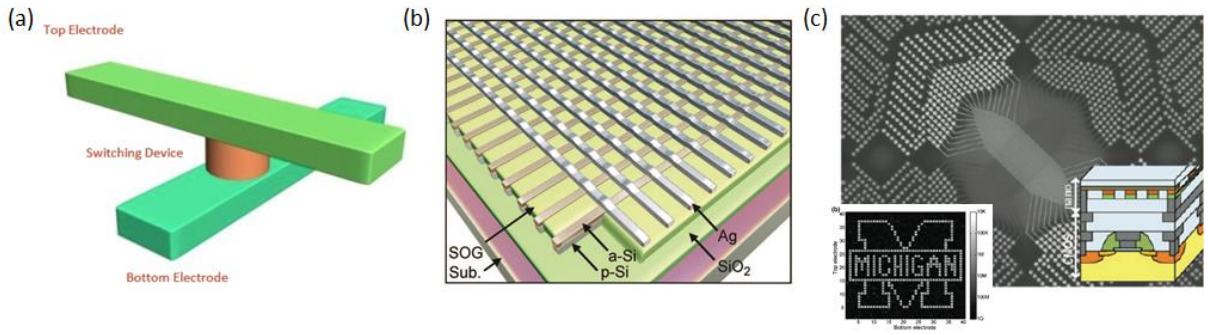

Figure 1.3: (a) Two-terminal structure of a RRAM device[8] (b) Crossbar array in which each RRAM cell is formed at every crosspoint of electrode lines. Reproduced with permission.[8] Copyright 2009, American Chemical Society. (c) SEM image of a RRAM crossbar array integrated on a top of CMOS chip. Reproduced with permission.[9] Copyright 2012, American Chemical Society.

scalability, RRAM is generally viewed as a promising candidate for a broad range of applications from high-density data storage, hybrid CMOS logic circuits to portable/wearable electronics.[9,10]

To further improve the storage density, three-dimensional(3D) integration of memory structures is necessary. 3D RRAM architectures can be achieved since only low-temperature processes are needed and the device does not rely on a single crystalline substrate. There are generally two types of 3D architectures.[11] The first involves stacking multiple 2D crossbar arrays on top of each other, as shown in Figure 1.4 (left), in which the storage density increases as the number of stacked layers increases. The fabrication and functionality of 3D stacked structures have been demonstrated using Ag-based ECM devices[12], polymer(PI:PCBM)-based organic resistive switching devices[13], and TaON-based VCM devices[14]. 3D stacking offers ultra-high scalability with minimal changes to the fabrication processes. However, the number of lithographic steps and masks used in the process increases linearly with the number of stacks.[11,15]

As a cost-effective 3D architecture, the vertical RRAM structure was proposed.[16] In the vertical RRAM structure the MIM cells are formed in parallel at the sidewall of a vertical electrode (Figure 1.4 right). Several studies have demonstrated the feasibility of the proposed vertical RRAM structure, mostly based on transition metal oxide-based systems *e.g.*, TaO_x[17], HfO_x[18],

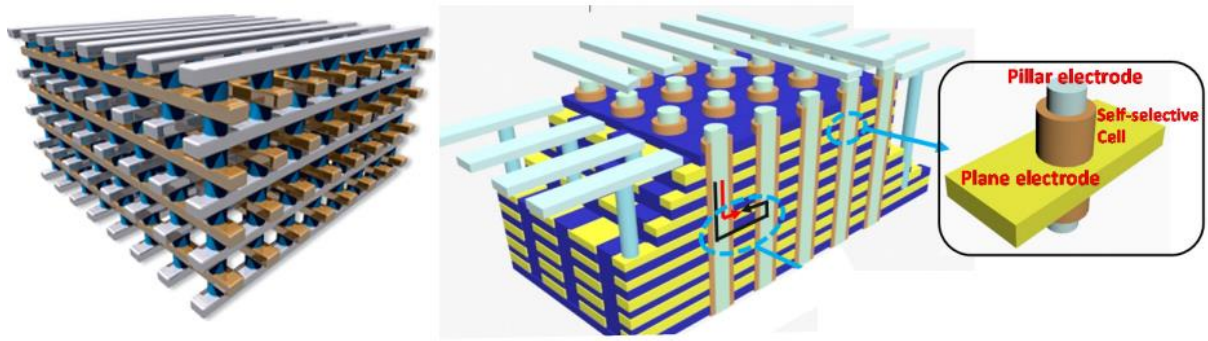

Figure 1.4: Schematic illustration of 3D X-point structure (left, Crossbar, Inc) and 3D vertical structure (right). Reproduced with permission.[22] Copyright 2012, IEEE.

TaO_x/TiO_2[19], WO_x[20], and $AlO_\delta/Ta_2O_{5-x}/TaO_y$[21]. The integrated RRAM devices in the vertical structure exhibit typical resistive switching behaviors, *e.g.*, >1000× on/off ratio, $>10^{10}$ endurance, 10^4 s retention at 125°C.[21] More importantly, good uniformity between devices at different layers was observed, supporting the feasibility of vertical 3D integration.[18,21] Since the electrode size in the vertical RRAM structure is determined by the width of the vertical electrode and the thickness of the horizontal electrode, potentially better scaling can be achieved since film thickness control during deposition is generally easier than pattern size control during lithography. For example, a graphene layer with atomic thickness has been employed as a horizontal electrode in a vertical RRAM structure, with improved power consumption due to the reduced device size.[23]

1.4 Emulation of Synaptic Plasticity by a RRAM Device

Beyond memory applications, an emerging, attractive field for resistance switching devices is bio-inspired neuromorphic computing.[24-26] Non-conventional computing architectures are now being seriously considered since the current digital computing architecture, in which the central processing unit (CPU) is physically separated from the memory, increasingly suffers from the von Neumann bottleneck problem, resulting in the loss of energy efficiency and throughput due to

intensive data movements. Furthermore, rapid advances in social networks, mobile devices and sensors require efficient, real-time storage and analysis of large amounts of data, while current hardware systems are not optimized for these data-intensive tasks. Neuromorphic computing systems, taking inspiration from the human brain, can potentially offer enormous computing capability through massive parallelism at extremely low power consumption, and have attracted strong interest as a promising option in the search of new computing paradigms.[27-32]

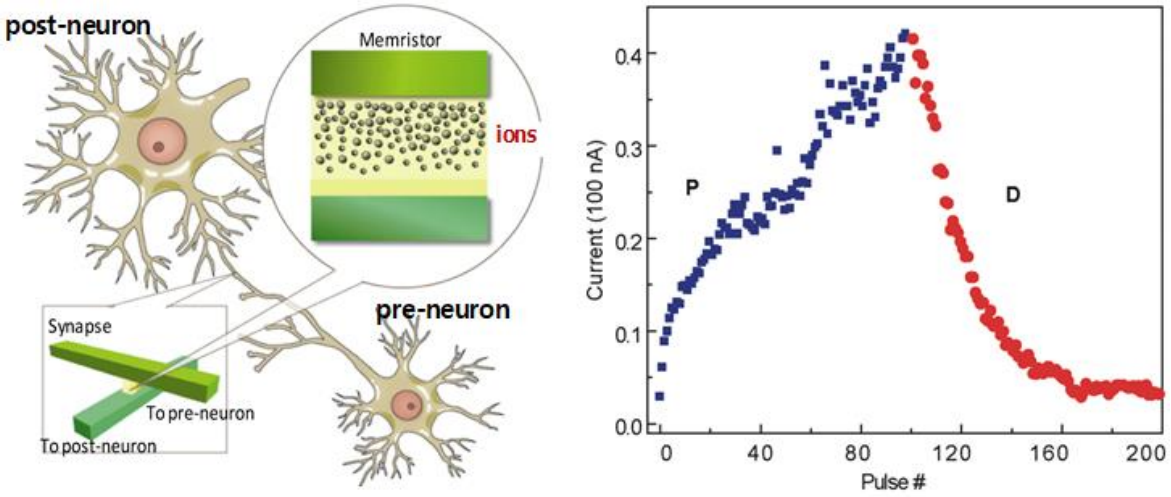

Figure 1.5: (a)Schematic illustration of a memristor acting as a synapse bridging a pair of neurons, with the ability to gradually modulate the synaptic weight (conductance) by controlling the internal ionic configuration. (b) Incremental conductance modulation of a memristor, in response to identical, consecutive potentiating (blue square) and depressing(red circle) pulses. Reproduced with permission.[24] Copyright 2010, American Chemical Society.

A biological network consists of interconnected neurons through reconfigurable connections – synapses, as illustrated in Figure 1.5a. The connection strength, the synaptic weight, can be strengthened (*via* potentiation) or weakened (*via* depression) depending on the neurons' firing patterns, *e.g.*, spiking timing and rates. The synapse can retain the updated weight[33] from tens of milliseconds to a few minutes in the form of short-term plasticity[34], and from minutes (and hours) to days (and months) in the form of long-term plasticity[35-37]. These effects, termed synaptic

plasticity, are believed to play a critical role in the processes of learning and memory.[38] The ability of RRAM devices to modulate its conductance and evolve into different states, dependent on the input history, makes these devices ideal building blocks as artificial synapses for hardware implementation of neuromorphic systems.[39-43] Additionally, analog-type switching, in which the device conductance changes incrementally in response to a series of input signals (spikes), allows simple implementation of online learning.[39,44] Such incremental conductance modulation effects, as shown in Figure 1.5b, were experimentally implemented in different types of memristive devices including WO_x, Ag/a-Si, Ta_2O_5 and Al_2O_3/TiO_2.[24,45-47]

1.5 Neuromorphic Computing based on RRAM Crossbar Network

An important feature of the cortical system is its large connectivity, enabling massively parallel processing. For example, the human brain has roughly 100 billion neurons with on average 7,000 synaptic connections per neuron[48]. The two-terminal structure of memristive devices can potentially offer high connectivity, high density, and random access that are required to implement

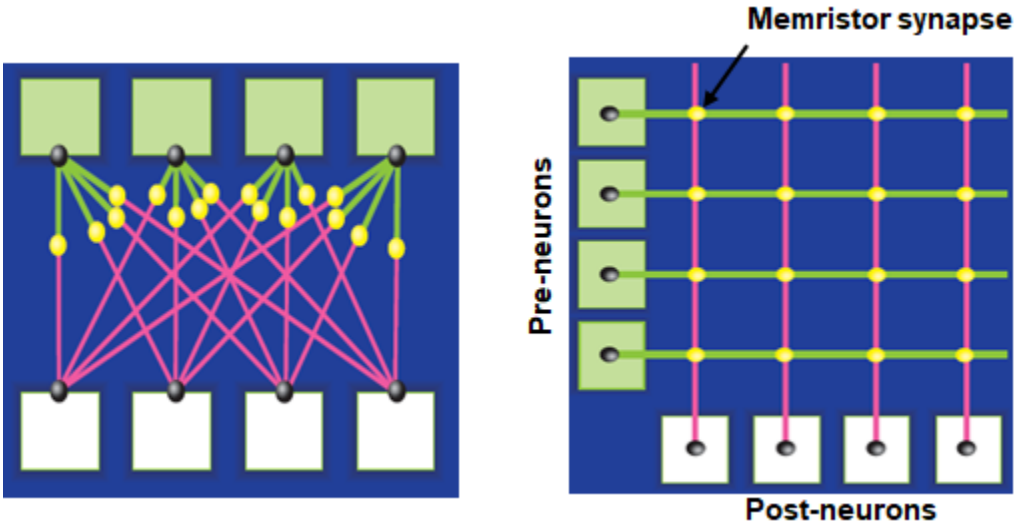

Figure 1.6: Schematic illustration showing how neural network can be mapped onto a crossbar structure in which RRAM devices formed at every crosspoint work as a synapse.

large-scale neuromorphic systems, in the form of RRAM crossbar arrays. Indeed, a neural network can be readily mapped onto a crossbar structure as shown in Figure 1.6. In this case, each input neuron (*e.g.* connected to a horizontal electrode) is connected to every output neuron (*e.g.* connected to a vertical electrode) with a memristor device acting as a synapse. Equally importantly, since current directly flows through the memristor between the input and output neurons, the "resistive" feature of the device allows it to directly modulate the information (in the form of current) flow through Ohm's law. In this sense, the device allows memory and computing functions to be performed simultaneously at the same physical location, without having to move data between different components.

Specifically, in a crossbar array, when input voltage pulses are applied to the rows of the crossbar, the output currents are determined by the product of the input voltage and the stored conductance values in the crossbar matrix, where the sum of currents from all devices connected

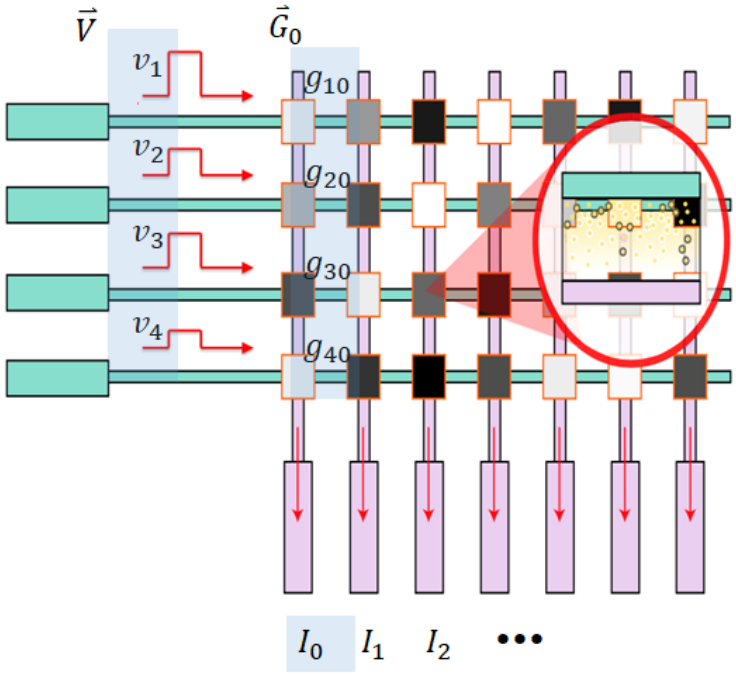

Figure 1.7: Schematic illustration of an artificial neural network based on a RRAM crossbar structure, in which vector-matrix multiplication operation can be parallel performed. Reproduced with permission.[49] Copyright 2017, Springer Nature.

in a column determines the total current collected at the column. In this way, the memristor crossbar naturally performs multiplication and accumulate (MAC) operations, while the output currents collected at the columns represent the (input) vector- (stored synaptic weight) matrix multiplication operation[40,43,47,50,51], a key operation in machine learning, analog computing, and other data intensive computing tasks. Note here the vector-matrix operation is achieved directly "for free" through physical phenomena (*e.g.* Ohm's law and Kirchhoff's law, where the output is obtained *via* a *single* read of the output current), without having to perform computation and move data between separate processor and weight storage units. The ability to change the device conductance *in situ* in turn allows the network to adapt to the input patterns and learn useful features from the input data using online learning algorithms,[40,41,43,47] thus making the crossbar-based system a natural fit for mapping machining-learning and bio-inspired computing algorithms.

1.6 Organization of the Thesis

This thesis discusses our work on developing analog switching RRAM device with improved incremental switching characteristics that are desirable for neuromorphic computing and possibly arithmetic computing systems. The studies are based on comprehensive understanding of the internal dynamics of resistive switching behaviors at atomic level, and implemented through systematic experimental efforts. The thesis is organized as following"

Chapter 1 is the introduction, where we provide the background information of RRAM device operation and applications, focusing on the concept of physical reconfiguration of nanomaterials through ionic transport and redox reactions. Potential of RRAM devices for high-density data storage and efficient parallel computing such as neuromorphic computing applications is introduced.

In Chapter 2, we investigate the electronic and optical properties of crystalline and amorphous Ta_2O_5 structures using first-principles calculations based on density functional theory (DFT) and the GW method. The calculated band gaps of the crystalline structures are too small to explain the experimental measurements, but the amorphous structure exhibits a strong exciton binding energy and an optical band gap (~ 4eV) in agreement with experiment. Additionally, short-range structure of our generated amorphous simulation cells shows excellent agreement with experimental data, allowing us to continue our theoretical calculations on a crucial defect, oxygen vacancy, in amorphous Ta_2O_5.

In Chapter 3, we investigate the thermodynamic, electronic, and optical properties of oxygen vacancies in amorphous Ta_2O_5 by first-principles calculations based on hybrid-functional DFT method. The calculated thermodynamic and optical transition levels are in good agreement with a broad range of diverse measured properties with various experimental methods, providing conclusive evidence for the identification of the defect states observed in experiments as originating from oxygen vacancies. Our results elucidate the fundamental atomistic properties of oxygen vacancies in various oxidation states as a function of growth conditions and provide guidance to control the properties of Ta_2O_5 films/devices.

In Chapter 4, we study the interactions between neutral and charged V_Os in amorphous Ta_2O_5 using DFT calculations. The binding energy between charged V_Os is strongly repulsive at short range, contradicting the experimentally observed high V_O concentration in filaments. On the other hand, neutral V_Os exhibit a short-range attraction that facilitates aggregation, but their charge neutrality precludes interactions with an electric field. We propose a series of charge-transition processes that are necessary during RS which enables V_O drift and aggregation. We experimentally support the proposed model with electrical measurements under visible-light illumination that

induce charge transitions of V_Os and enhance the programming and erasing processes. Our results provide microscopic understanding of the RS mechanism and the effect of light on ions/defects migration.

In Chapter 5, we show that the RS characteristics in tantalum-oxide-based memristors can be systematically tuned by inserting a graphene film with engineered nanopores. Graphene, with its atomic thickness and excellent impermeability and chemical stability, can be effectively integrated into the device stack and offer unprecedented capabilities for the control of ionic dynamics at nanoscale. In this device structure, the graphene film effectively blocks ionic transport and redox reactions, thereby the oxygen vacancies required during the RS process are only allowed to transport through the engineered nanosized openings in the graphene layer, leading to effective modulation of the device performance by controlling the nanopore size in graphene. The roles of graphene as an ion-blocking layer in the device structure were further supported by transmission electron microscopy (TEM), energy-dispersive X-ray spectroscopy (EDS) and atomistic simulations based on first-principles calculations.

In Chapter 6, we report experimental implementation of memristor crossbar hardware systems using the optimized devices. The system we developed can allow direct comparison of the Euclidean distances without normalizing the weights. Unsupervised *K*-means clustering algorithm through online learning was experimentally implemented. The approaches and devices can be used in other unsupervised learning systems, and significantly broadens the range of problems memristor based network can solve.

1.7 References

1. Valov, I. & Lu, W. D. Nanoscale electrochemistry using dielectric thin films as solid electrolytes. Nanoscale 8, 13828–13837 (2016).

2. Lee, J. & Lu, W. D. On-Demand Reconfiguration of Nanomaterials: When Electronics Meets Ionics. Advanced Materials 30, 1702770–33 (2017).

3. Celano, U. et al. Imaging the Three-Dimensional Conductive Channel in Filamentary-Based Oxide Resistive Switching Memory. Nano Letters 15, 7970 (2015).

4. Waser, R., Dittmann, R., Staikov, G. & Szot, K. Redox-Based Resistive Switching Memories – Nanoionic Mechanisms, Prospects, and Challenges. Advanced Materials 21, 2632–2663 (2009).

5. Chua, L. Memristor-The missing circuit element. IEEE Transactions on Circuit Theory 18, 507–519 (1971).

6. Kang, S. M. & Chua, L. Memristive devices and systems. Proc. IEEE 64, 209–223 (1976).

7. Strukov, D. B., Snider, G. S., Stewart, D. R. & Williams, R. S. The missing memristor found. Nature 453, 80–83 (2008).

8. Jo, S. H., Kim, K.-H. & Lu, W. High-Density Crossbar Arrays Based on a Si Memristive System. Nano Letters 9, 870–874 (2009).

9. Kim, K.-H. et al. A Functional Hybrid Memristor Crossbar-Array/CMOS System for Data Storage and Neuromorphic Applications. Nano Letters 12, 389–395 (2012).

10. Jo, S. H. & Lu, W. CMOS Compatible Nanoscale Nonvolatile Resistance Switching Memory. Nano Letters 8, 392–397 (2008).

11. Edwards, A. H. et al. Reconfigurable Memristive Device Technologies. Proc. IEEE 103, 1004–1033 (2015).

12. Kügeler, C., Meier, M., Rosezin, R., Gilles, S. & Waser, R. High density 3D memory architecture based on the resistive switching effect. Solid State Electronics 53, 1287–1292 (2009).

13. Song, S. et al. Three-Dimensional Integration of Organic Resistive Memory Devices. Advanced Materials 22, 5048–5052 (2010).

14. Hsieh, M.-C. et al. Ultra high density 3D via RRAM in pure 28nm CMOS process. in 10.3.1–10.3.4 (IEEE, 2013). doi:10.1109/iedm.2013.6724600

15. Seok, J. Y. et al. A Review of Three-Dimensional Resistive Switching Cross-Bar Array Memories from the Integration and Materials Property Points of View. Adv. Funct. Mater. 24, 5316–5339 (2014).

16. Yoon, H. S. et al. Vertical cross-point resistance change memory for ultra-high density non-volatile memory applications. in 26–27 (IEEE, 2009).

17. Baek, I. G. et al. Realization of vertical resistive memory (VRRAM) using cost effective 3D process. in 31.8.1–31.8.4 (IEEE, 2011). doi:10.1109/iedm.2011.6131654

18. Chen, H.-Y. et al. HfOx based vertical resistive random access memory for cost-effective 3D cross-point architecture without cell selector. in 20.7.1–20.7.4 (IEEE, 2012). doi:10.1109/iedm.2012.6479083

19. Hsu, C.-W. et al. 3D vertical TaO<inf>x</inf>/TiO<inf>2</inf> RRAM with over 10³ self-rectifying ratio and sub-μA operating current. in 10.4.1–10.4.4 (IEEE, 2013). doi:10.1109/iedm.2013.6724601

20. Gaba, S., Sheridan, P., Du, C. & Lu, W. 3-D Vertical Dual-Layer Oxide Memristive Devices. IEEE Trans. Electron Devices 61, 2581–2583 (2014).

21. Bai, Y. et al. Study of Multi-level Characteristics for 3D Vertical Resistive Switching Memory. Scientific Reports 4, 3213 (2014).

22. Luo, Q. et al. 8-Layers 3D vertical RRAM with excellent scalability towards storage class memory applications. in 2.7.1–2.7.4 (IEEE, 2017). doi:10.1109/iedm.2017.8268315

23. Lee, S., Sohn, J., Jiang, Z., Chen, H.-Y. & Wong, H.-S. P. Metal oxide-resistive memory using graphene-edge electrodes. Nature Communications 8407 (2015). doi:10.1038/ncomms9407

24. Jo, S. H. et al. Nanoscale Memristor Device as Synapse in Neuromorphic Systems. Nano Letters 10, 1297–1301 (2010).

25. Ohno, T. et al. Short-term plasticity and long-term potentiation mimicked in single inorganic synapses. Nature Materials 10, 591–595 (2011).

26. Pickett, M. D., Medeiros-Ribeiro, G. & Williams, R. S. A scalable neuristor built with Mott memristors. Nature Materials 12, 114–117 (2012).

27. Pershin, Y. V. & Di Ventra, M. Neuromorphic, Digital, and Quantum Computation With Memory Circuit Elements. Proc. IEEE 100, 2071–2080 (2012).

28. Modha, D. S. et al. Cognitive computing. Communications of the ACM 54, 62 (2011).

29. Indiveri, G. et al. Neuromorphic Silicon Neuron Circuits. Frontiers in Neuroscience 5, (2011).

30. Poon, C.-S. & Zhou, K. Neuromorphic Silicon Neurons and Large-Scale Neural Networks: Challenges and Opportunities. Frontiers in Neuroscience 5, (2011).

31. Mead, C. Neuromorphic electronic systems. Proc. IEEE 78, 1629–1636 (1990).

32. Indiveri, G. & Liu, S.-C. Memory and Information Processing in Neuromorphic Systems. Proc. IEEE 103, 1379–1397 (2015).

33. Cowan, W. M., Südhof, T. C. & Stevens, C. F. Synapses. (2001).

34. Zucker, R. S. & Regehr, W. G. Short-Term Synaptic Plasticity. Annual Review of Physiology 64, 355–405 (2002).

35. Bliss, T. V. P. & Collingridge, G. L. A synaptic model of memory: long-term potentiation in the hippocampus. Nature 361, 31–39 (1993).

36. Abraham, W. C. How long will long-term potentiation last? Philosophical Transactions of the Royal Society B: Biological Sciences 358, 735–744 (2003).

37. Bliss, T. V. P. & Lømo, T. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. The Journal of Physiology 232, 331–356 (1973).

38. Sjöström, P. J., Turrigiano, G. G. & Nelson, S. B. Rate, Timing, and Cooperativity Jointly Determine Cortical Synaptic Plasticity. Neuron 32, 1149–1164 (2001).

39. Yang, J. J., Strukov, D. B. & Stewart, D. R. Memristive devices for computing. Nature Nanotechnology 8, 13–24 (2013).

40. Chang, T., Yang, Y. & Lu, W. Building Neuromorphic Circuits with Memristive Devices. IEEE Circuits and Systems Magazine 13, 56 (2013).

41. Kuzum, D., Yu, S. & Wong, H.-S. P. Synaptic electronics: materials, devices and applications. Nanotechnology 24, 382001 (2013).

42. Jeong, D. S., Kim, K. M., Kim, S., Choi, B. J. & Hwang, C. S. Memristors for Energy-Efficient New Computing Paradigms. Adv. Electron. Mater. 2, 1600090 (2016).

43. Sheridan, P. & Lu, W. Memristors and Memristive Devices for Neuromorphic Computing. Memristor Networks 129–149 (2014). doi:10.1007/978-3-319-02630-5_8

44. Sheridan, P. M., Du, C. & Lu, W. D. Feature Extraction Using Memristor Networks. IEEE Transactions on Neural Networks and Learning Systems 27, 2327–2336 (2016).

45. Chang, T. et al. Synaptic behaviors and modeling of a metal oxide memristive device. Appl. Phys. A 102, 857–863 (2011).

46. Kim, S., Choi, S., Lee, J. & Lu, W. D. Tuning Resistive Switching Characteristics of Tantalum Oxide Memristors through Si Doping. ACS Nano 8, 10262–10269 (2014).

47. Prezioso, M. et al. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. Nature 521, 61–64 (2015).

48. Drachman, D. A. Do we have brain to spare? Neurology 64, 2004–2005 (2005).

49. Sheridan, P. M. et al. Sparse coding with memristor networks. Nature Nanotechnology 12, 784–789 (2017).

50. Gao, L., Alibart, F. & Strukov, D. B. Analog-input analog-weight dot-product operation with Ag/a-Si/Pt memristive devices. in 88–93 (IEEE, 2012). doi:10.1109/VLSI-SoC.2012.7332082

51. Merced-Grafals, E. J., Davila, N., Ge, N., Williams, R. S. & Strachan, J. P. Repeatable, accurate, and high speed multi-level programming of memristor 1T1R arrays for power efficient analog computing applications. Nanotechnology 27, 365202 (2016).

Chapter 2 Electronic Band Structure of Tantalum Pentoxide Polymorphs

2.1 Introduction

As discussed in Chapter 1, RRAM device operations are based on ionic transport and electrochemical reactions, leading to the formation/rupture of conductive filaments at the nanoscale. However, atomic-level thermodynamic and kinetic descriptions of resistive switching in oxide-based RRAM, which are critical for continued device design and optimization, are still relatively lacking. Comprehensive understanding and optimizations of RRAM devices will require the development of atomistic descriptions of the fundamental properties of the switching materials and the mobile defects.

Tantalum oxide-based RRAM devices have been found to exhibit high endurance, long retention, and fast switching speed.[1-3] Despite the application potential of this material, consistent experimental and theoretical data about its structural, electronic, and optical properties are still missing, which is a major obstacle to further improve the material performance in applications.

In this work[4], we systematically investigate the electronic and optical properties of three crystalline (β-Ta_2O_5, δ-Ta_2O_5, and 11 f.u. model) and four amorphous structures of tantalum pentoxide using first-principles calculations based on DFT and many-body perturbation theory. We employed the GW approximation[5] to obtain quasiparticle band structures for all examined Ta_2O_5 structures without adjustable or empirical parameters to uncover the relationship between the underlying structure and the resulting electronic and optical properties. Our results show that

the amorphous structure has an optical gap that is in agreement with experiment and is thus a plausible structural model for low-temperature Ta_2O_5 (L-Ta_2O_5).

2.2 Discrepancy in Previous Experiments and Theoretical Calculations

Various crystalline polymorphs for L-Ta_2O_5 have been proposed in the literature. A model consisting of 11 formula units (11 f.u. model) was proposed based on x-ray powder diffraction measurements.[6] Subsequently, two simpler models, the orthorhombic β-Ta_2O_5 and the hexagonal δ-Ta_2O_5, were reported.[7,8] Previous results for the electronic properties of these three polytypes at the density functional theory (DFT)/generalized gradient approximation (GGA) level are consistent with each other, but results obtained by more advanced methods (i.e., the HSE06 and PBE0 hybrid functionals and DFT+U calculations) show large variations. Gu et al. calculated band gaps of 0.2 and 1.06 eV for β-Ta_2O_5 and δ-Ta_2O_5, respectively, at the GGA level.[9] Wu et al. found that the band gap using GGA (HSE06) is 0.1 eV (0.9 eV) for β-Ta_2O_5 and 1.1 eV (2 eV) for δ-Ta_2O_5.[10] Nashed et al. reported gaps of 0.2 eV (2.45 eV) for β-Ta_2O_5 and 1.04 eV (2.92 eV) for δ-Ta_2O_5 with the PBE (PBE0) functional.[11] GGA+U^d+U^p calculations by Kim et al. found band gaps of 2.24 and 3.24 eV for β- and δ-Ta_2O_5, respectively.[12] Moreover, there is a large discrepancy between calculated and experimental values of the band gap. Experimental measurements on the band gap of tantalum pentoxide grown by different deposition techniques find consistent values around 4 eV,[13] which is much larger than calculated values for β- and δ-Ta_2O_5 (Table 2.1).

More recently, structural models with lower symmetry have been examined to resolve this crystalline structure and band-gap issue. Even though the 11 f.u. model[14] of L-Ta_2O_5 is the structure reported earliest in the literature, it has not been computationally studied in detail due to

Table 2.1: The calculated band gap of tantalum pentoxide polymorphs (eV).

	Method	β-Ta_2O_5[a]	δ-Ta_2O_5[b]	11 f.u.[b]	λ-Ta_2O_5[a]	Pseudo-hexagonal	amorphous
This work	PBE	0.19	1.27	1.92	2.18	·	2.98
	GW	1.03	2.22 (2.79)[c]	2.96 (3.59)[c]	3.23	·	4.26 (3.92)[d]
Reference 14	GGA	0.2	1.06	·	·	·	·
Reference 15	GGA	0.1	1.1	·	·	·	·
	HSE06	0.9	2	·	·	·	·
Reference 16	PBE	0.2	1.04	·	·	·	·
	PBE0	2.45	2.92	·	·	·	·
	B3LYP	2.15	2.65	·	·	·	·
Reference 21	HSE06	~ 2	~ 2	·	~ 4	·	~ 4
Reference 20	PBE	·	·	1.96	·	·	·
	PBE0	·	·	3.91	·	·	·
Reference 17	GGA+U^d+U^p	2.24	3.24	4.8	4.7	4.28	·

[a] Direct band gap.
[b] Indirect band gap.
[c] Minimum direct band gap.
[d] Optical gap including excitonic effects.

its large unit cell and complicated atomic configuration. Recently, it was found to be more stable and to have a larger gap than the higher-symmetry models.[14] Yang et al. found its gap to be 3.91 eV within PBE0,[14] while Kim et al. reported a band gap of 4.8 eV using the GGA+U^d+U^p method[12]. Furthermore, other proposed crystalline models are found to be stable and exhibit band gaps similar to experiment. The λ-Ta_2O_5 polytype with a symmetry of *Pbam* proposed by Lee et al. has a band gap of ~4 eV and lower total energy than the β-Ta_2O_5, δ-Ta_2O_5, and 11 f.u. models as determined with HSE06.[15] In addition, Kim et al. generated a pseudo-hexagonal Ta_2O_5 structure that is modified from the high-symmetry δ-Ta_2O_5 geometry by breaking the symmetry and optimizing the atomic positions, and has a band gap of 4.28 eV as determined within GGA+U^d+U^p.[12]

However, the reported gap values for the various Ta_2O_5 polytypes determined by more advanced methods are not in agreement with each other. The GGA+U^d+U^p calculation by Kim et al. for λ-Ta_2O_5 found a band gap of 4.7 eV,[12] while this value is ~4 eV with HSE06.[15] The

calculated gap values for β-Ta_2O_5, δ-Ta_2O_5, and 11 f.u. models within GGA+U^d+U^p are larger by approximately 1 eV compared to hybrid-functional results.[10,14] Moreover, the calculated band gap of β-Ta_2O_5 by Lee et al. using HSE06 is larger by about 1 eV compared to the HSE06 result of another study.[10] In light of the wide variation and inconsistency of these results it is impossible to reliably relate the underlying polytype geometry to the band-gap value and to establish a connection with electronic and optical measurements.

2.3 Methodology

The charge density and electronic wave functions were obtained with DFT in the GGA[16] for the exchange-correlation potential using the Vienna ab initio simulation package (VASP).[17,18] We used projector augmented-wave (PAW)[19,20] pseudopotentials with a plane-wave cutoff of 400 eV and included the Ta $5s^2 5p^6 5d^3 6s^2$ and O $2s^2 2p^4$ electrons in the valence. The quasiparticle band structures were calculated using the one-shot GW method within VASP. The GW method has been validated to accurately predict the electronic properties of bulk materials[21-23] and nanostructures[24-26]. The maximally localized Wannier function formalism was employed to interpolate the band energies to arbitrary points in the first Brillouin zone. [27] The exciton binding energies and optical absorption spectra were determined using the Bethe-Salpeter equation (BSE) method and the k-point interpolation scheme implemented in BerkeleyGW. [28] Quasiparticle corrections to the DFT eigenvalues for the BSE calculations were applied with a scissor-shift operator fitted to the GW results of VASP.

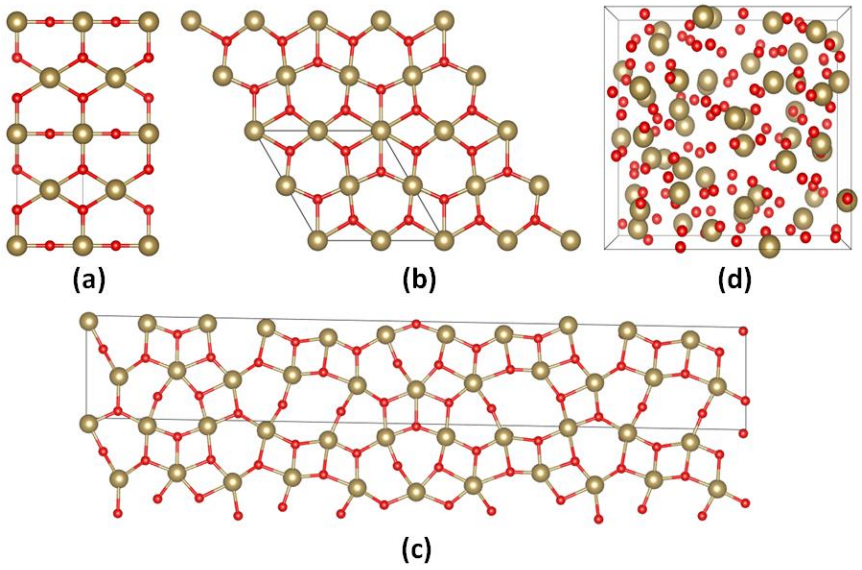

Figure 2.1: Crystal structures of (a) β-Ta_2O_5, (b) δ-Ta_2O_5, (c) the 11 f.u. model, and (d) amorphous tantalum pentoxide. Reproduced with permission.[4] Copyright 2014, AIP Publishing.

2.4 Band Structure of Ta_2O_5 Polymorphs

We calculated the band gaps of the β-Ta_2O_5, δ-Ta_2O_5, 11 f.u. model crystalline structures (Figure 2.1) but the results are too small to explain experiment. The β-Ta_2O_5 polytype has an orthorhombic cell with *Pccm* symmetry containing four tantalum and ten oxygen atoms. Experimental values for the lattice parameters and atomic positions of β-Ta_2O_5 were used for the calculations.[7] The PBE band structure of β-Ta_2O_5 has a 0.19 eV direct band gap at Z, in good agreement with previous calculations.[10,11] The GW band structure of β-Ta_2O_5 is shown in Figure 2.2a. GW widens the gap to 1.03 eV, in agreement with previous HSE06 calculations.[10] However, the value of the gap for this polytype is too small compared to the experimental value of ~4 eV.[13] The δ-Ta_2O_5 structure belongs to the *P6/mmm* group and contains two formula units in one unit cell. The GW band structure of δ-Ta_2O_5 exhibits a 2.22 eV indirect band gap and a 2.79 eV direct band gap, Figure 2.2b, in agreement with previous HSE06 calculations[10] but much smaller than experiment. The 11 f.u. model has an orthorhombic unit cell with 22 tantalum and 55 oxygen

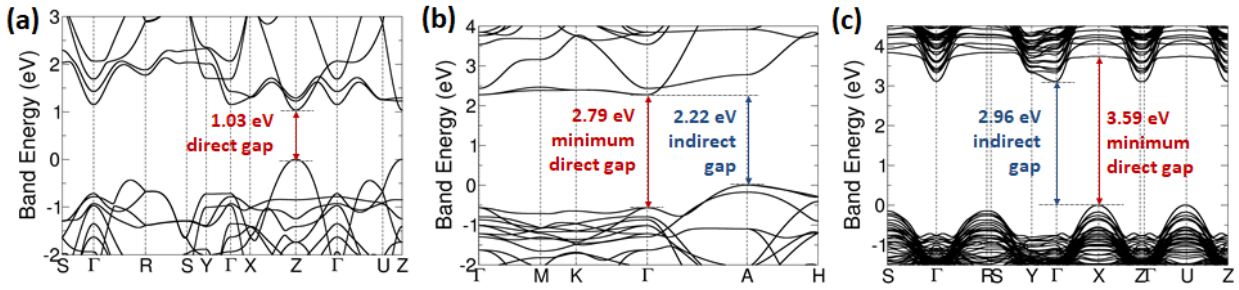

Figure 2.2: Band structures of (a) β-Ta2O5, (b) δ-Ta2O5, and (c) the 11 f.u. model of tantalum pentoxide calculated with the GW method. β-Ta2O5 has a direct band gap of 1.03 eV at Z, while δ-Ta2O5 has an indirect band gap of 2.22 eV between A and Γ with a minimum direct band gap of 2.79 eV. The band gap of the 11 f.u. model is indirect with a magnitude of 2.96 eV, while the minimum direct gap is 3.59 eV. Reproduced with permission.[4] Copyright 2014, AIP Publishing.

atoms. In our calculation the occupancy of the oxygen sites is determined by the combination that gives the lowest total energy as reported by Yang et al.[14] The atomic positions and cell parameters of the 11 f.u. model are optimized with a structural relaxation calculation. The GW band structure of the 11 f.u. model displays a 2.96 eV indirect band gap (Figure 2.2c). The minimum direct band gap of this model is 3.59 eV, which is larger than the β-Ta_2O_5 and δ-Ta_2O_5 structures, but it is ~0.4 eV smaller than the experimental optical gap.

On the other hand, we calculated the optical gap of amorphous Ta_2O_5 in good agreement with experiment. The amorphous structures were generated using first-principles molecular-dynamics simulations with the VASP code[29] starting from the crystalline structure of β-Ta_2O_5 with a supercell contacting 24 Ta_2O_5 formula units. The Brillouin zone was sampled at the Γ point and the density was set to the experimental value of 6.88 g/cm^3. [30] We annealed the cell at a temperature of 3500 K using a Nosé-Hoover thermostat[31] followed by a quenching step down to 1000 K below the theoretical melting point. [32] Finally, a damped molecular-dynamics run was performed to optimize the structural geometry. We carried out the melt and quench procedure with various annealing times (12 ps, 16 ps, and 20 ps) and cooling rates (300 K/ps and 200 K/ps) to generate

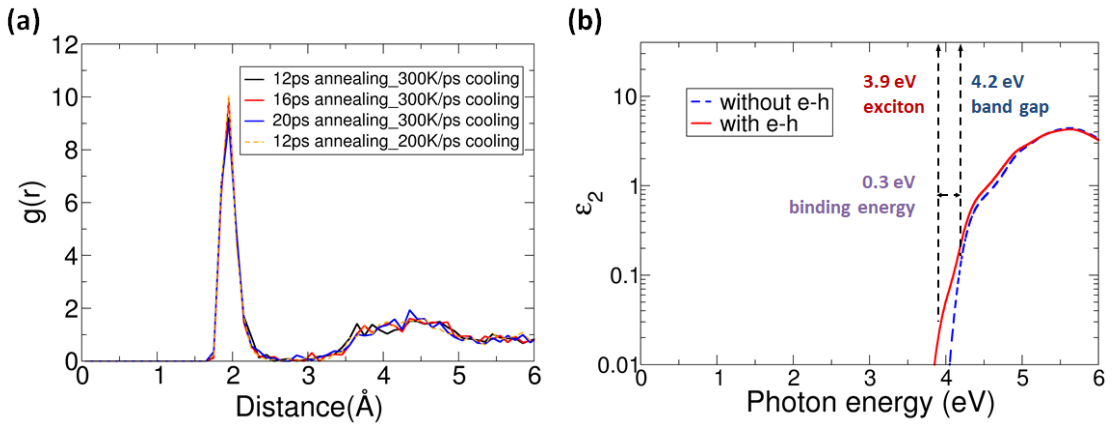

Figure 2.3: (a)Radial distribution function of the amorphous Ta_2O_5 structures generated with different annealing times and cooling rates. (b)Imaginary part of the dielectric function ε_2 for amorphous Ta_2O_5 with and without electron-hole interaction effects included. The exciton binding energy (0.3 eV) was obtained from the difference between the optical gap at 3.92 eV (i.e., the lowest exciton energy) and the electronic band gap at 4.26 eV. Reproduced with permission.[4] Copyright 2014, AIP Publishing.

four amorphous cells under different annealing and cooling conditions. The four structures have consistent atomic geometries as shown in the radial distribution function of Ta-O (Figure 2.3a) with similar short-range structure as experimental and previous Monte-Carlo computational results.[10,33] The GW gaps of the different amorphous structures are also consistent with each other. The average of the GW band gaps is 4.26 eV with a standard deviation of 0.1 eV. We also performed BSE calculations including the electron-hole interaction, which gives rise to excitonic effects that lower the energy of the first optical excitation. Figure 2.3b shows the calculated imaginary part of the dielectric function including excitonic effects. The optical calculation gives the lowest exciton energy of 3.92 eV, which is in agreement with the experimental optical gap. The difference between the lowest exciton energy and the electronic band gap is the exciton binding energy of 0.3 eV. The size of the exciton was determined by projecting the square of the exciton electron wave function along one of the simulation cell directions. The full-width at half-maximum (FWHM) of the envelope function is 21.5 Å.

2.5 Relationship Between the Electronic Properties vs. Underlying Structure

To analyze the dependence of the band gap on the atomic geometry we determined the atomic orbitals that form the valence band maximum (VBM) and conduction band minimum (CBM) of each polymorph. As expected, the VBM consist of oxygen 2*p* orbitals and the CBM consists primarily of tantalum 5*d* orbitals in all polymorphs. Interestingly, it was found that the small-gap structures, i.e., β-Ta_2O_5 and δ-Ta_2O_5, have different CBM orbital configurations than the larger-gap 11 f.u. model and amorphous structures. There is a significant tantalum 6*s* component in the CBM wave functions of β-Ta_2O_5 and δ-Ta_2O_5, but not the larger-gap structures. In addition, this tantalum 6*s* orbital is attributed to specific Ta atoms in the two models. For example, there are two kinds of Ta atom sites in β-Ta_2O_5, (2a) and (2b) as shown in Figure 2.4a. The tantalum 6*s* orbital that contributes to the CBM wave function comes only from the Ta atoms in the (2b) site. Figure 2.4 shows the total density of states (DOS), as well as the projected density of states (P-DOS) obtained by projecting onto Wannier functions localized on each Ta atom. Specific atoms such as

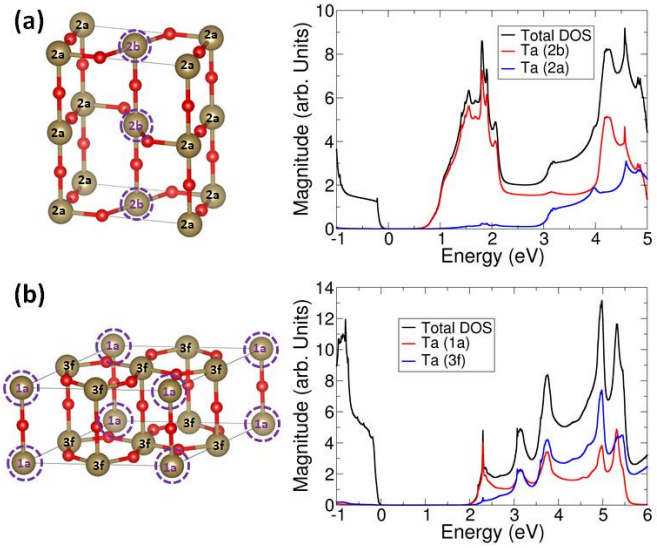

Figure 2.4: Crystal structure, density of states (DOS), and projected density of states (P-DOS) on Wannier functions localized on each Ta atom for (a) β-Ta_2O_5 and (b) δ-Ta_2O_5. Reproduced with permission.[4] Copyright 2014, AIP Publishing.

the Ta atoms at (2b) site of β-Ta_2O_5 and the Ta atom at the (1a) site of δ-Ta_2O_5 contribute to the DOS near the band edge.

Our results establish the connection between the underlying structure and the electronic properties of Ta_2O_5. It is expected that the small-gap structures would show larger band gap if the contribution by the (2b) Ta atoms in β-Ta_2O_5 and the (1a) Ta atoms of δ-Ta_2O_5 was removed. Indeed, this hypothesis is supported by the recent results of Helali et al., who reported an increased band gap for the reconstructed hexagonal δ′-Ta_2O_5 structure. [34] The band-gap increase is related to ionic displacements of the (1a) Ta and the nearby O ions in the reconstructed δ′-Ta_2O_5 structure, which reduces the coordination of the (1a) Ta atoms from 8-fold to 6- or 7-fold and increases the band gap. A similar increase of the band gap by reconstruction was observed in the pseudo-hexagonal Ta_2O_5 structure by Kim et al.[12]

2.6 Conclusion

In summary, we studied the electronic and optical properties of crystalline and amorphous Ta_2O_5 with first-principles calculations. The calculated band gaps of the examined crystalline structures (β-Ta_2O_5, δ-Ta_2O_5, and 11 f.u. models) are too small to explain the experimental measurements, but the amorphous structure gives an optical band gap that agrees with experimental results. Our results therefore provide evidence that the amorphous structure is a plausible structural model for the L-Ta_2O_5 materials with a measured optical gap of ~4 eV as reported in the literature.

2.7 References

1. Lee, M.-J. *et al.* A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O5−x/TaO2−x bilayer structures. *Nature Materials* **10,** 625–630 (2011).

2. Yang, Y., Sheridan, P. & Lu, W. Complementary resistive switching in tantalum oxide-based resistive memory devices. *Applied Physics Letters* **100,** 203112–5 (2012).

3. Kim, S., Choi, S., Lee, J. & Lu, W. D. Tuning Resistive Switching Characteristics of Tantalum Oxide Memristors through Si Doping. *ACS Nano* **8,** 10262–10269 (2014).

4. Lee, J., Lu, W. & Kioupakis, E. Electronic properties of tantalum pentoxide polymorphs from first-principles calculations. *Applied Physics Letters* **105,** 202108 (2014).

5. Hybertsen, M. S. & Louie, S. G. Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies. *Phys. Rev. B* **34,** 5390–5413 (1986).

6. Stephenson, N. C. & Roth, R. S. Structural systematics in the binary system Ta2O5–WO3. V. The structure of the low-temperature form of tantalum oxide L-Ta2O5. *Acta Crystallogr B Struct Crystallogr Cryst Chem* **27,** 1037–1044 (1971).

7. Aleshina, L. A. & Loginova, S. V. Rietveld analysis of X-ray diffraction pattern from β-Ta2O5 oxide. *Crystallography Reports* **47,** 415–419 (2002).

8. Fukumoto, A. & Miwa, K. Prediction of hexagonalTa2O5structure by first-principles calculations. *Phys. Rev. B* **55,** 11155–11160 (1997).

9. Gu, T., Wang, Z., Tada, T. & Watanabe, S. First-principles simulations on bulk Ta2O5 and Cu/Ta2O5/Pt heterojunction: Electronic structures and transport properties. *Journal of Applied Physics* **106,** 103713–9 (2009).

10. Wu, Y.-N., Li, L. & Cheng, H.-P. First-principles studies of Ta 2O 5polymorphs. *Phys. Rev. B* **83,** 485–9 (2011).

11. Nashed, R., Hassan, W. M. I., Ismail, Y. & Allam, N. K. Unravelling the interplay of crystal structure and electronic band structure of tantalum oxide (Ta 2O 5). *Phys. Chem. Chem. Phys.* **15,** 1352–1357 (2013).

12. Kim, J.-Y. *et al.* Electronic structure and stability of low symmetry Ta 2O 5polymorphs. *phys. stat. sol. (RRL)* **8,** 560–565 (2014).

13. Fleming, R. M. *et al.* Defect dominated charge transport in amorphous Ta2O5 thin films. *Journal of Applied Physics* **88,** 850–862 (2000).

14. Yang, Y., Nahm, H.-H., Sugino, O. & Ohno, T. Electronic structures of oxygen-deficient Ta 2O 5. *AIP Advances* **3,** 042101–9 (2013).

15. Lee, S.-H., Kim, J., Kim, S.-J., Kim, S. & Park, G.-S. Hidden Structural Order in Orthorhombic Ta2O5. *Phys. Rev. Lett.* **110,** 235502–5 (2013).

16. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Phys. Rev. Lett.* **78,** 1396–1396 (1997).

17. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47,** 558–561 (1993).

18. Kresse, G. & Furthmüller, J. Efficient iterative schemes forab initiototal-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54,** 11169–11186 (1996).

19. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50,** 17953–17979 (1994).

20. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59,** 1758–1775 (1999).

21. Shi, G. & Kioupakis, E. Relativistic quasiparticle band structures of Mg 2Si, Mg 2Ge, and Mg 2Sn: Consistent parameterization and prediction of Seebeck coefficients. *Journal of Applied Physics* **123,** 085114–13 (2018).

22. Bayerl, D. & Kioupakis, E. Theoretical limits of thermoelectric figure of merit in n-type TiO2polymorphs. *Phys. Rev. B* **91,** 165104–5 (2015).

23. Mengle, K. A., Shi, G., Bayerl, D. & Kioupakis, E. First-principles calculations of the near-edge optical properties of β-Ga2O3. *Applied Physics Letters* **109,** 212104–5 (2016).

24. Sanders, N., Bayerl, D., Shi, G., Mengle, K. A. & Kioupakis, E. Electronic and Optical Properties of Two-Dimensional GaN from First-Principles. *Nano Letters* **17,** 7345–7349 (2017).

25. Bayerl, D. *et al.* Deep ultraviolet emission from ultra-thin GaN/AlN heterostructures. *Applied Physics Letters* **109,** 241102–6 (2016).

26. Toulouse, A. S. *et al.* Frenkel-like Wannier-Mott excitons in few-layer PbI2. *Phys. Rev. B* **91,** 808–5 (2015).

27. Marzari, N., Mostofi, A. A., Yates, J. R., Souza, I. & Vanderbilt, D. Maximally localized Wannier functions: Theory and applications. *Rev. Mod. Phys.* **84,** 1419–1475 (2012).

28. Deslippe, J. *et al.* BerkeleyGW: A massively parallel computer package for the calculation of the quasiparticle and optical properties of materials and nanostructures. *Computer Physics Communications* **183,** 1269–1289 (2012).

29. Kresse, G. & Hafner, J. Ab initiomolecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B* **49,** 14251–14269 (1994).

30. Kakio, S., Hosaka, K., Arakawa, M., Ohashi, Y. & Kushibiki, J.-I. Surface Acoustic Wave Properties of Amorphous Ta$_{2}$O$_{5}$ and Nb$_{2}$O$_{5}$ Thin Films Prepared by Radio Frequency Sputtering. *Jpn. J. Appl. Phys.* **51,** 07GA01–6 (2012).

31. Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **31,** 1695–1697 (1985).

32. Garg, S. P., Krishnamurthy, N., Awasthi, A. & Venkatraman, M. The O-Ta (Oxygen-Tantalum) system. *Journal of Phase Equilibria* **17,** 63–77 (1996).

33. Kimura, H., Mizuki, J., Kamiyama, S. & Suzuki, H. Extended x-ray absorption fine structure analysis of the difference in local structure of tantalum oxide capacitor films produced by various annealing methods. *Applied Physics Letters* **66,** 2209–2211 (1995).

34. Helali, Z., Calatayud, M. & Minot, C. Novel Delta-Ta 2O 5Structure Obtained from DFT Calculations. *The Journal of Physical Chemistry C* **118,** 13652–13658 (2014).

Chapter 3 Electronic and Optical Properties of Oxygen Vacancy in Amorphous Tantalum Oxide

3.1 Introduction

In oxide-based RRAM devices including tantalum oxide, oxygen vacancies (V_O) are commonly considered as crucial defects that can significantly influence the electrical and optical properties of Ta_2O_5 films and modulate the operation of the devices. Specifically, the operation of memristive devices essentially relies on the control of oxygen vacancies, where the resistance switching is triggered by the formation, dissolution, and redistribution of oxygen vacancies followed by the formation or rupture of conductive filaments.[1-3] Despite extensive experimental studies on oxygen vacancies in Ta_2O_5, the reported oxygen vacancy defect states have been ambiguous, lacking accurate and conclusive theoretical evidence.

In this work[4], we use first-principles computational methods based on hybrid density functional theory (DFT) to investigate the thermodynamic, electronic, and optical properties of oxygen vacancies in amorphous Ta_2O_5. Amorphous cells that were proven to exhibit good agreement with experimental data in structural and electronic properties as discussed in Chapter 2 were employed to study the oxygen vacancy properties in this work. We show that thermodynamically preferred oxygen sites with lower vacancy formation energies are those that primarily contribute to the valence band maximum wave function. We also determine the concentrations and the dominant oxidation states of oxygen vacancies as a function of growth conditions. Our results on the electronic and optical properties indicate that experimentally

observed but as-yet unidentified defect states in Ta_2O_5 can be attributed to neutral oxygen vacancies. By analysing the charge distributions and energies of excess electrons in the presence of oxygen vacancies we observe the formation of polarons bound to charged oxygen vacancies, which transport via a hopping conduction mechanism and can have a significant effect on the films' electrical conductivity. Our work provides microscopic understanding of the properties of oxygen vacancies in Ta_2O_5, which is important for tuning its electrical and optical properties, and for optimizing material growth and device operation.

3.2 Previous Experimental and Theoretical Studies on Oxygen Vacancy in Ta_2O_5

Due to their significant role in device operations, oxygen vacancies in Ta_2O_5 have been extensively studied both experimentally and theoretically. A broad range of values for defect levels in Ta_2O_5 have been reported from various experimental methods, including thermally stimulated current (TSC),[5] fitting of *I-V* measurements,[6] photoconductivity,[7,8] and photoluminescence.[9] The defects observed in the experimental studies have been tentatively identified as oxygen vacancies, despite the absence of unequivocal theoretical or experimental evidence. Lau *et al*. found a defect state at 0.8 eV below the conduction band by the TSC method.[5] In addition, Mikhaelashvili *et al*. reported a defect ionization energy of 0.58 eV extracted from the fitting of *I-V* measurement to the equation of the Poole-Frenkel conduction mechanism (*i.e.*, the transport of electrons via repeated trapping and detrapping by defect levels).[6] They tentatively attributed the observed defect states to oxygen vacancies. Moreover, the photon energy for the onset of photocurrent, at which defect-trapped electrons get photoexcited into the conduction band and conduct, was reported at 1.5 eV, without an apparent photoconductivity peak.[7,8] In addition, Markosyan *et al*. reported intrinsic and induced optical absorption losses in amorphous Ta_2O_5 films at a photon energy of 1.165 eV using a dual-beam experiment, [10] which they assign to absorption introduced from point defects

such as oxygen vacancies or oxygen interstitials. However, the nature of defect states around 1.1 eV from the band edges of Ta_2O_5 is unknown. On the theoretical side, early density functional theory (DFT) studies using the local density approximation (LDA) and the generalized gradient approximation (GGA) to the exchange-correlation functional do not fully explain the defect levels measured in experiment because the values of the band gaps of insulators and the positions of defect states are not accurate with these two functionals.[11,12] Other studies did not make a distinction between thermodynamic and optical transitions, and they also drew conclusions by comparing experimental data to inaccurate LDA calculations.[5] In addition, most theoretical studies on defects in Ta_2O_5 employed the crystalline structure. However, according to our previous theoretical work employing the GW method, the calculated band gaps of the purported crystalline structures are too small to explain experimental values, while the calculated optical gap of the amorphous structure is in good agreement with experiment.[13] Recent theoretical studies have focused on explaining the resistive-switching mechanism in Ta_2O_5 memristors.[14-16] For example, Jiang and Stewart used first-principles nudged-elastic-band calculations and reported a low energy barrier for oxygen vacancy migration, which is attributed to the adaptive crystal structure and accounts for the low operating voltage and fast switching speed of Ta_2O_5-based memristive devices.[16] Moreover, Xiao et al. studied oxygen vacancies in amorphous Ta_2O_5 in the context of the mechanism of resistive switching.[14] However, their calculated value for the band gap (3.3 eV)[14] is noticeably smaller than experimentally measured values (~ 4 eV)[7]. Moreover, the reported defect level (0.8 eV below the conduction band) determined by their density-of-states calculation[14] is related to optical transitions, not the thermodynamic transition level reported experimentally[5]. Calculations of thermodynamic transition levels require total-energy differences of defects at different charge states that take into account the structural relaxation of nearby atoms, whereas

transition levels from the density of states (which assume atoms remain unrelaxed) should be compared to optical-transition levels. Overall, theoretical studies to comprehensively characterize and understand the fundamental properties of oxygen vacancies in Ta_2O_5 (such as their charge state, formation energy, charge distribution, and thermodynamic and optical transition energies) that explain the wide range of available experimental data are still missing.

3.3 Methodology

We performed first-principles calculations based on density functional theory (DFT), which is well suited for the study of defects in periodic simulation cells. [17] The equilibrium geometry of the atoms, the total energies, and electronic wave functions are obtained using a plane-wave pseudopotential method within the generalized gradient approximation (GGA) functional of Perdew, Burke, and Ernzerhof (PBE).[18] Projector augmented-wave (PAW) pseudopotentials[19,20], as implemented in the Vienna ab-initio simulation package (VASP),[21,22] were used with a plane-wave cutoff of 400 eV. A k-point mesh of 1 x 1 x 1 was used to ensure convergence of 5 meV/atom for the total energy and 20 meV/defect for the defect formation energy. Atomic configuration of cells with vacancy was obtained through force minimization where the force tolerance for geometry relaxation was 0.02 eV/Å. The range-separated hybrid functional of Heyd, Scuseria, and Ernzerhof (HSE06)[23] was used to describe the exchange-correlation interaction of electrons in electronic-structure and geometric-relaxation calculations, which yields more accurate values for the band gap than PBE. A mixing ratio of 0.2 and a range separation parameter of 0.2 $Å^{-1}$ are used in all calculations.

The formation energy of oxygen vacancies for various charge states is calculated by:

$$E^f\left[V_O^q\right] = E_{tot}\left[\mathrm{Ta_2O_5} : V_O^q\right] - E_{tot}[\mathrm{Ta_2O_5} : \text{bulk}] + \mu_O + qE_F \qquad (3.1)$$

where $E_{tot}[\mathrm{Ta_2O_5} : V_O^q]$ is the total energy of the amorphous Ta_2O_5 cell containing an oxygen vacancy with charge q, and $E_{tot}[\mathrm{Ta_2O_5} : \mathrm{bulk}]$ is the total energy for the perfect cell. The parameters μ_O and E_F are the chemical potential of an oxygen atom and the Fermi level of electrons. The Fermi level is referenced to the valence band maximum (VBM) in the bulk cell with the correction of band alignments.[24] We also evaluated the finite-size correction for charged defects in supercell calculations.[25,26] The correction for the long-range Coulomb interaction is quite small, ~ 0.01 eV, due to the high dielectric constant of Ta_2O_5 of 24, and it does not affect our conclusions.

3.4 Formation Energy of Oxygen Vacancies in Amorphous-Ta_2O_5

We performed first-principles calculations based on hybrid-functional density functional theory (DFT) to study oxygen vacancies in amorphous Ta_2O_5 using a simulation cell that contains 24 Ta and 60 O atoms generated by the melt-and-quench method (Figure 3.1a).[13] The calculated band gap of the amorphous cell used in this work is 4.24 eV based on the HSE method, which is consistent with previous experimental and theoretical results in considering the exciton binding energy of 0.3 eV.[7,13] Seventeen of the 60 possible oxygen sites with different atomic configurations, indicated in Figure 3.1a, were chosen and the formation energy of oxygen vacancies for various charge states was calculated to identify the thermodynamically preferred sites. We found that the formation energy of V_O does not correlate to geometrical properties of the oxygen sites such as their coordination number or the bond lengths with nearest Ta ions. However, we uncovered a correlation between the vacancy formation energy and the amount of the partial charge density for the valence band maximum (VBM) wave function of the corresponding oxygen atom. Oxygen sites that contribute the most to the VBM partial charge density exhibit lower

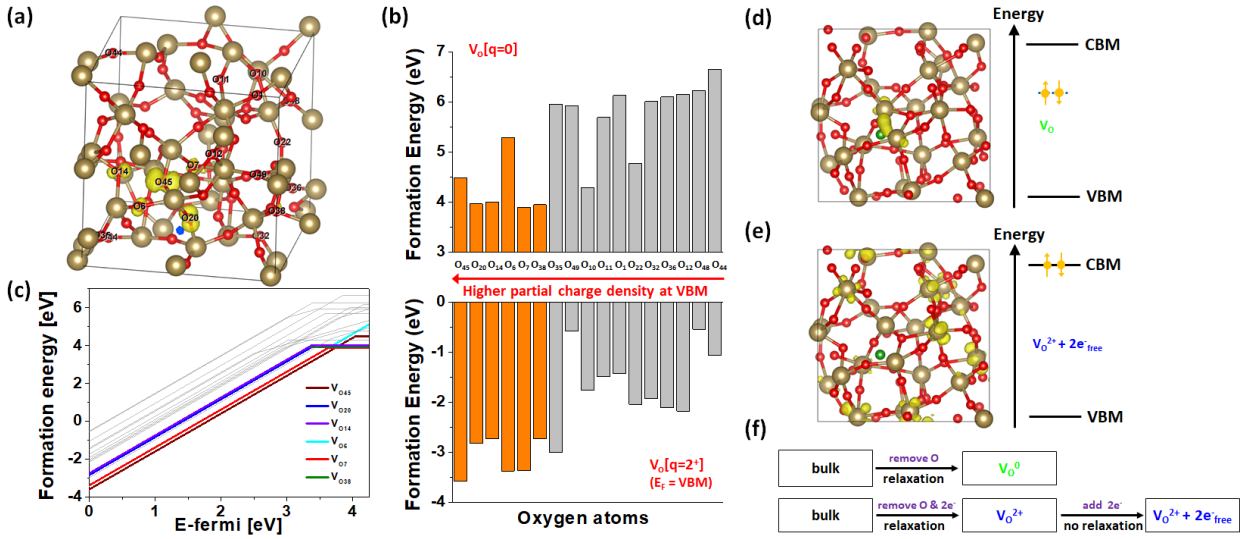

Figure 3.1: (a) Atomic structure of the amorphous Ta_2O_5 simulation cell and isosurface of the band-decomposed charge density at the VBM. Gold (large) and red (small) spheres represent tantalum and oxygen atoms, respectively. The value of the isosurface is chosen to enclose 50% of the electron density. (b) Formation energies of neutral (top) and doubly charged (bottom) oxygen vacancies in order of increasing contribution of the corresponding O atom to the partial charge density at the VBM. The orange-shaded values denote oxygen sites with significant partial charge density (larger than 2.5 %) and the grey values are oxygen sites with negligible partial charge density (less than 0.5%). (c) Formation energies of oxygen vacancies as a function of Fermi level for O-rich conditions. Oxygen vacancies at O sites with larger amounts of partial charge density at the VBM (indicated by colored lines) have lower formation energy relative to vacancies at O sites with negligible partial charge density (grey lines). The Fermi level is referenced to the VBM and its maximum limit is the CBM. The calculated band gap using the HSE06 functional is 4.24 eV. (d)-(e) Charge distributions of excess electrons and schematic of the energy position of defect states and electron occupancy in cells with (d) neutral and (e) ionized oxygen vacancies. Neutral oxygen vacancies give rise to occupied mid-gap electronic states that are localized between two neighbor Ta ions. The two electrons from fully ionized oxygen vacancies behave as free delocalized electrons. (f) The computational procedure of generating the V_O defect configurations to study neutral (d) and ionized (e) oxygen vacancies. Reproduced with permission.[4] Copyright 2017, Royal Society of Chemistry.

vacancy formation energies, and all oxygen sites having non-negligible contributions to the VBM partial charge density are included in the seventeen oxygen sites investigated in this work. This allows us to find the most favorable oxygen sites for the formation of vacancies without having to consider all oxygen sites that requires substantial computational resources.

The wave function squared of the VBM state is concentrated around O_{45}, as shown in Figure 3.1a. The O_{45} and neighbouring oxygen ions, O_{20}, O_{14}, O_6, O_7, and O_{38}, whose *2p* orbitals

primarily contribute to the VBM wave function, exhibit lower formation energies than other oxygen ions with lower contributions to the VBM. For example, the average value of the formation energy for neutral oxygen vacancies (V_O^0) is 5.80 eV for the oxygen sites that contribute less than 0.5% VBM partial charge density and 4.27 eV for the oxygen sites with larger (more than 2.5%) VBM partial charge (Figure 3.1b). To determine stable oxygen vacancies in various charge states, the formation energy of the oxygen vacancies is plotted as a function of the Fermi-level in Figure 3.1c. The correlation between thermodynamic stability and the amount of the VBM partial charge density is observed over the entire range of the Fermi level within the gap. We attribute this correlation to the lower energy required to extract valence electrons occupying states near the VBM than deeper electrons. This correlation can also provide useful guidance to identify thermodynamically favoured defect sites in large and complex structures of other amorphous materials as well. Moreover, the thermodynamically most stable oxygen vacancies show negative-U behaviour, *i.e.,* the singly charged vacancy (V_O^+) state is thermodynamically unstable. Similar negative-U behaviour is observed for oxygen vacancies in ZnO. [27] In addition, the transition level between the +2 and 0 charge states for the most stable O vacancies is located near the conduction band edge, thus oxygen vacancies in amorphous Ta_2O_5 behave as donors.

We also investigate the charge distributions and electronic levels of excess electrons introduced by oxygen vacancies. The band-decomposed charge density for the excess electrons in the presence of neutral and ionized oxygen vacancies is plotted in Figures 3.1d and e, respectively. The excess electrons in cells with neutral oxygen vacancies are localized in the empty space between Ta ions and give rise to a fully occupied defect level within the forbidden gap (Figure 3.1d). On the other hand, no states are observed within the gap for the Vo^{2+} case, and electrons of ionized oxygen vacancies behave as free electrons delocalized over many Ta ions (Figure 3.1e).

3.5 Equilibrium Oxygen Vacancy Concentration

To determine the equilibrium oxygen vacancy concentration at room temperature and their dominant charge state, the equilibrium Fermi level due to the interplay of neutral and ionized oxygen vacancies was calculated self consistently as a function of chemical potential. On the one hand, the formation energy of charged vacancies depends on the Fermi level as in Figure 3.1c, but on the other hand the Fermi energy is controlled by the concentration of extra electrons released into the conduction band by charged defects. In the case of amorphous Ta_2O_5, the formation energy of V_O^{2+} is negative when the Fermi level is located near the valence band edge, which favours the spontaneous generation of charged oxygen vacancies.

However, the electrons released into the conduction band by the formation of charged vacancies shift the Fermi level to higher values, increase the charged-defect formation energy, and inhibit the further spontaneous generation of charged vacancies. By self-consistently evaluating the feedback between the Fermi-level position within the gap, the vacancy formation energy, and

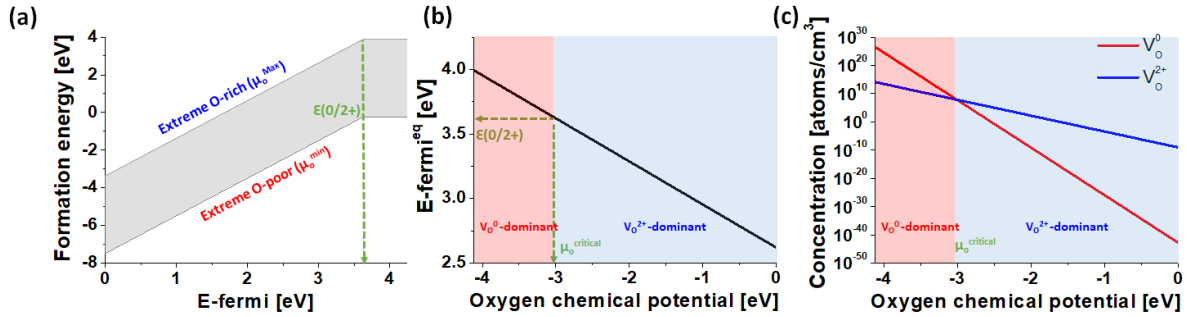

Figure 3.2: (a) Formation energy of an oxygen vacancy at a low-formation-energy O site (V_{O7}) as a function of Fermi-level within the range of growth conditions. The grey area indicates the range of formation energies between the extreme O-rich and O-poor limits. (b) Calculated equilibrium Fermi level as a function of oxygen chemical potential. The dominant oxidation state of the oxygen vacancy depends on the value of the chemical potential. (c) Calculated equilibrium concentration of neutral and doubly charged oxygen vacancies as a function of the oxygen chemical potential. Only neutral oxygen vacancies can occur in large concentrations in equilibrium, and are thus the dominant O vacancy charge state in experimentally grown samples. Reproduced with permission.[4] Copyright 2017, Royal Society of Chemistry.

the concentration of oxygen vacancies, we determine the equilibrium position of the Fermi level and the vacancy concentrations. Moreover, the vacancy formation energy depends on the growth conditions through the oxygen chemical potential, μ_O. Figure 3.2a exhibits the formation energy of an oxygen vacancy at a characteristic oxygen site, V_{O7}, in both neutral and doubly charged states as a function of μ_O within the range of possible values. The vacancy formation energies and hence the equilibrium Fermi level depend on the μ_O. The equilibrium Fermi level increases towards the conduction band as μ_O decreases, as shown in Figure 3.2b. This is because lower values of μ_O (oxygen-poor conditions) lower the ionized oxygen vacancy formation energies and increase the concentration of excess electrons in the conduction band. The concentration of vacancies at different charge states and the dominant charge state are affected by the growth conditions related with the chemical potentials because the equilibrium Fermi level changes, as shown in Figure 3.2b. There is a critical value for the oxygen chemical potential μ_O^{critical} below which neutral oxygen vacancies are thermodynamically preferred (O-poor conditions), while for μ_O above the critical value (O-rich conditions) charged vacancies dominate. At the critical μ_O value, the equilibrium Fermi level is equal to the thermodynamic transition level for which the dominant oxidation state changes (Figure 3.2b and c). As seen in Figure 3.2c, under O-poor conditions the neutral vacancy concentrations reach large values and dominate compared to charged vacancies. In contrast, the concentration of charged vacancies under O-rich conditions is low (lower than 10^{10} cm^{-3} at room temperature). Therefore, our results demonstrate that thermodynamically stable oxygen vacancies in amorphous Ta_2O_5 predominantly exist in the neutral charge state, and thus neutral oxygen vacancies with low formation energies are a possible cause of mid-gap defect levels.

3.6 Thermodynamic Transition Level

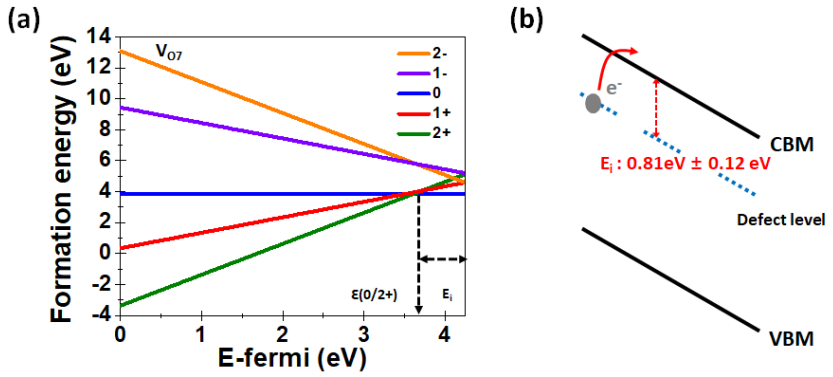

Figure 3.3: (a) Calculated formation energies for various oxidation states of O vacancies as a function of Fermi level for one of the low-formation-energy O sites (V_{O7}) under O-rich conditions. The transition level, ε(0/2+) is 0.61 eV below the bottom of conduction band, which corresponds to the ionization energy of the oxygen vacancy. (b) Schematic of the Poole-Frenkel conduction mechanism in insulators. Electrons transport by getting trapped and detrapped between different defect sites and the conduction band. The average calculated value of the ionization energy for stable oxygen vacancies, indicated by the red arrow, is 0.81 eV ± 0.12 eV, which is in good agreement with experimental values of 0.6–0.8eV. Reproduced with permission.[4] Copyright 2017, Royal Society of Chemistry.

Figure 3.3a shows the formation energy of a stable oxygen vacancy (V_{O7}) for various oxidation states as a function of Fermi level. Negatively charged (corresponding to the case of additional electrons trapped at the already occupied defect sites) and positively single-charged vacancies (V_O^n where n=–1, –2, and +1) are not stable over the entire range of Fermi levels within the gap. The thermodynamic transition level, i.e. the position of the Fermi level for which the formation energies of the +2 and 0 oxidation states are equal, for V_{O7} is 0.61 eV below the conduction band minimum (CBM). Positively double-charged oxygen vacancies (V_O^{2+}) are thermodynamically stable below this value, while neutral oxygen vacancies (V_O^0) are more stable above the transition level. Experimentally, the thermal ionization energy of defects corresponds to this thermodynamic transition level.[28] We determined the thermodynamic transition level, ε(0/2+) for four stable oxygen vacancies with the lowest neutral-vacancy formation energies (V_{O7}, V_{O14}, V_{O20}, and V_{O38}. The average value of the transition level for these four vacancies is 0.81 eV with

a standard deviation of 0.12 eV (Figure 3.3b), which is in agreement with the measured values of 0.8 eV by TSC measurements,[5] and 0.58 eV by fitting to the Poole-Frenkel conduction mechanism.[6] The agreement of the calculated thermodynamic transition level with the experimental ionization energy presents conclusive evidence that the experimentally observed defect states at an energy of 0.6 – 0.8 eV below the conduction band minimum can be identified with oxygen vacancies. On the other hand, the large value of the ionization energy (0.81 eV) is too large to be overcome with thermal energy at room temperature. Our results therefore indicate that isolated oxygen vacancies in Ta_2O_5 are unlikely to be ionized at room temperature and hence act as deep donors.

3.7 Optical Transition Level

To determine the position of defect levels and interpret experimental data, we need to invoke the configuration coordinate diagram (Figure 3.4) and distinguish between thermodynamic and optical transitions due to their different time scales. Since charged defects in different oxidation states have distinct charge distributions and equilibrium atomic lattice configurations, transitions between those oxidation states may cause the rearrangement of the atoms around the defect site. For thermodynamic transitions, there is sufficient time for atoms to relax to the atomic geometry of the new ground state. On the contrary, optical transitions occur at a much shorter timescale and atoms are assumed to remain frozen at the geometry of the initial state during the transition.[28] Moreover, the rearrangement of the atoms between the geometries of the initial and final defect oxidation states causes the Stokes shift of the optical absorption and emission energies. Figure 3.4a represents the calculated configuration coordinate diagram of V_{O7} in the neutral, +1, and +2 charge states. Electrons in the charged states are assumed to be located at the CBM. Optical

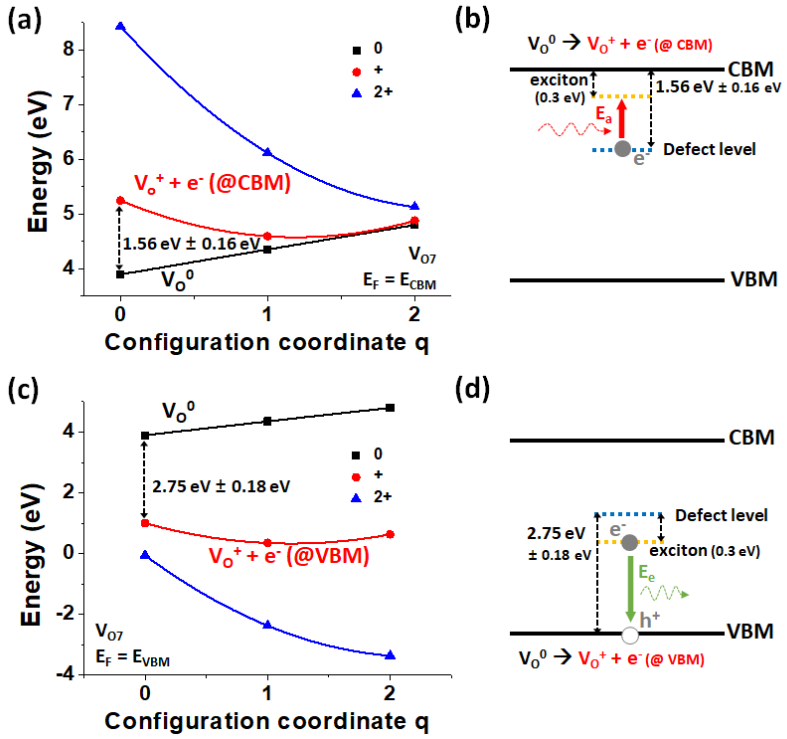

Figure 3.4: (a) Calculated configuration coordinate diagram for one oxygen vacancy (V_{O7}) in the neutral, +1, and +2 charge states. The Fermi level is assumed to be located at the bottom of the conduction band. The average calculated value of the lowest-energy optical transition for neutral defects is 1.56 eV. (b) Schematic of the defect state induced by neutral oxygen vacancies within the gap. The red arrow indicates the optical transition that excites trapped electrons at mid-gap defect states to the conduction band by light absorption. Considering the strong electron-hole attraction and exciton binding energy in amorphous Ta_2O_5 (0.3 eV) the absorption energy (E_a) is estimated to be 1.26 eV. (c) Calculated configuration coordinate diagram of V_{O7} assuming the Fermi level to be located at the valence band maximum. The average value of the neutral O vacancy mid-gap levels for stable oxygen vacancies is 2.75 eV above the VBM. (b) Energy diagram of mid-gap state induced by neutral oxygen vacancies. The green arrow indicates the radiative recombination of electrons trapped at defect states with holes in the valence band. The emitted photon energy (E_e) considering the exciton binding energy in amorphous Ta_2O_5 (0.3 eV) is estimated to be 2.45 eV. Reproduced with permission.[4] Copyright 2017, Royal Society of Chemistry.

transition energies are determined by fixing the atomic configuration to that of the initial oxidation state. For instance, the photon energy needed to excite an electron from the neutral vacancy to the CBM is calculated using the equilibrium configuration of the initial neutral vacancy state, and corresponds to the vertical energy difference between the 0 and +1 charge states at the atomic configuration of the neutral state (Figure 3.4a and b). The average calculated value of the

photoexcitation energy for the four most stable neutral oxygen vacancies is 1.56 eV with a standard deviation of 0.16 eV, which is similar to the onset of the experimental photoconductivity data.[7,8] However, unlike photoconductivity experiments, electron-hole interactions need to be considered to determine the transition energies of defect states in optical-absorption experiments. [29] Electron-hole interactions lower the transition energy by an amount equal to the exciton binding. According to our previous work, the exciton binding energy in amorphous Ta_2O_5 is 0.3 eV.[13] Assuming that bulk and defect-state excitons have similar binding energies, we find that electron-hole interactions lower the optical absorption energy from 1.56 eV to 1.26 eV (Figure 3.4b). Therefore, the reported optical absorption at 1.165 eV in experiment[10] can be ascribed to excitations from the defect state of neutral oxygen vacancies at 1.56 eV below the CBM to the lowest bound exciton state. Similarly, the light-emission energy associated with neutral oxygen vacancies can be obtained from the configuration-coordinate diagram. After optical excitation, the atoms around an oxygen vacancy site relax to a new equilibrium configuration and emit radiation with a longer wavelength than the absorbed one as shown in Figure 3.4a. The average value of the emission for the optical transition from V_O^+ (E_F=E^{CBM}) to V_O^0 is 0.33 eV with a Stokes shift of 1.2 eV. However, photoluminescence of 0.3 eV has not been reported experimentally for Ta_2O_5. Nonetheless, Figure 3.4c also indicates the $V_O^0 \rightarrow V_O^+$ (E_F=E^{VBM}) radiative recombination of electrons at the defect level with holes in the valence band. Our results predict that photoluminescence for this transition occurs at 2.45 eV (including the exciton binding energy of 0.3 eV, as in Figure 3.4d), which is in reasonable agreement with the experimentally reported luminescence at 2.2 eV.[9]

Our results about the effect of growth conditions on the stability of oxygen vacancies in each oxidation state provide new insights about the structure-property relations of Ta_2O_5. By adjusting the growth conditions, such as the oxygen partial pressure to tune the oxygen chemical

potential and the addition of other dopants to adjust the Fermi level, the dominant charge state of oxygen vacancies (+2 or neutral) in Ta_2O_5 can be controlled. This enables the tuning of the electrical and optical properties of Ta_2O_5, since charged vacancies introduce n-type doping and no sub-band-gap optical absorption, while neutral vacancies introduce deep occupied states in the gap as well as optical absorption in the visible. Moreover, atomistic phenomena that involve the migration and the aggregation/dissolution of oxygen vacancies, such as resistive switching in memristors, can be interpreted and analysed in terms of transitions between different charge states of oxygen vacancies due to changes in the local environment during the switching process.

3.8 Formation of Polaron

Our calculations also hint to the formation of polarons bound to charged oxygen vacancies in Ta_2O_5. We first calculate the polaron formation energy in bulk (defect-free) Ta_2O_5. The polaronic configuration has lower energy than the delocalized configuration by 58 meV (Figure 3.5a). The extra electron becomes localized on a few Ta sites after relaxation (Figure 3.5) and causes a local lattice distortion: the nearest-neighbour O ions move outward from the Ta ions due to repulsion by the excess electrons and the bond length increases by ~1%.

We subsequently examine polaronic effects in oxygen vacancies. By adding two electrons into previously relaxed V_O^{2+} simulation cells and further relaxing (as described in Figure 3.6b) we investigate the relative thermodynamic stability of three possible electron configurations: (a) the neutral vacancy configuration, with electrons bound to mid-gap defect states (Figure 3.1d), (b) the fully charged vacancy, with electrons occupying delocalized conduction-band states (Figure 3.1e), or (c) a bound polaron state, which involves electrons occupying spatially localized states in the vicinity of charged vacancies. Most thermodynamically preferred vacancy sites reverted to the

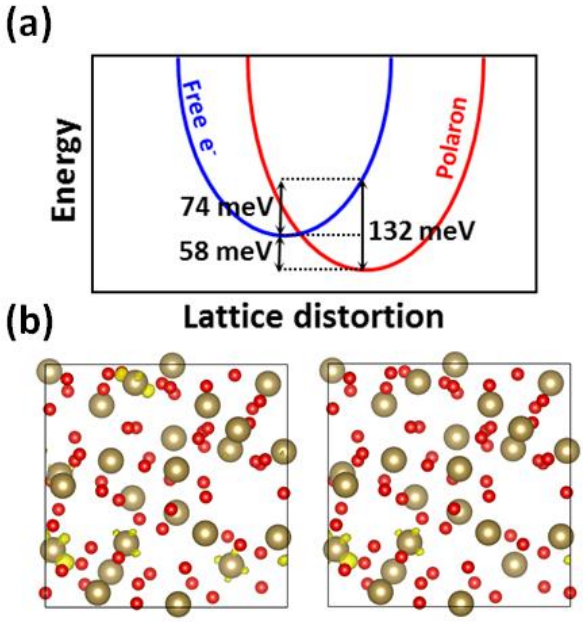

Figure 3.5: (a) Calculated configuration coordinate diagram for a delocalized electron and a polaron in amorphous Ta_2O_5. The ground-state configuration of the polaron has lower energy by 58 meV than the delocalized electron. (b) Charge distributions of an extra electron in unrelaxed (left) and relaxed (right) cells. The charge distribution of excess electron becomes more localized after relaxation, giving rise to the formation of polarons. The value of the isosurface was chosen to enclose 10% of the electron density in both cases. Reproduced with permission.[4] Copyright 2017, Royal Society of Chemistry.

neutral-vacancy configuration (a) after the addition of two electrons and the relaxation step. However, certain oxygen sites (V_{O7}, and V_{O45}) also exhibit a polaronic state (c) at a higher energy (0.3 eV on average) than the neutral one (a). We also found one particular vacancy site (V_{O6}) for which the lowest-energy state is the polaronic one, (c), although its energy is 0.3 eV on average higher than the lowest-energy neutral state (a) of other thermodynamically preferred sites. After accounting for all vacancies, polarons, and their complexes we expect that a fraction of electrons that are bound to the most stable neutral-vacancy states will overcome this 0.3 eV average energy difference and form polaron states at high temperature. In all cases, bound complexes between polarons and charged vacancies (c) have lower formation energies than fully ionized vacancies and free electrons (b). For the polaronic states the two excess electrons occupy two distinct same-spin eigenstates inside the forbidden gap, which results in a triplet system with two unpaired

electrons. Also, the electrons are neither localized at the vacancy sites nor delocalized over the entire crystal, but they are instead localized at individual Ta sites forming polarons, as shown in Figure 3.6a. Our calculated value of the binding energy of the vacancy–polaron complexes is 0.2 eV, which is similar to theoretical values for vacancy-polaron binding energies in TiO_2.[30]

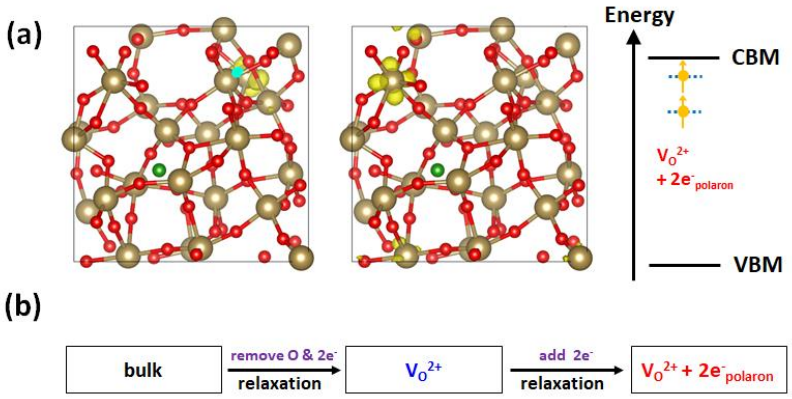

Figure 3.6: (a) Charge distribution of excess electrons and schematic of the position of defect states and electron occupancy for an O vacancy site (V_{O45}) that leads to the formation of spin-polarized polarons. The two electrons are neither delocalized nor localized at the oxygen vacancy site, but instead are localized on Ta ions and form polarons. The two electrons in the $V_O^{2+} + 2e^-_{polaron}$ case occupy two different eigenstates with the same spin, resulting in a triplet state. (b) The computational procedure of generating V_O configurations that give rise to the $V_O^{2+} + 2e^-_{polaron}$ case. Two electrons are added to the V_O^{2+} cell and subsequently the atoms are relaxed to find the new energy minimum. Reproduced with permission.[4] Copyright 2017, Royal Society of Chemistry.

The formation of polarons greatly affects the electron mobility and conduction mechanism as reported in experimental and theoretical research on TiO_2[30,31] since polarons transport by a hopping mechanism. Even though the existence of polarons in Ta_2O_5 has been postulated based on experimental data showing hopping-transport behaviour, [32,33] it has not been reported and analysed by theoretical work. Our results show that oxygen vacancies act as deep donors in amorphous Ta_2O_5 and that most carriers are localized at neutral vacancies, but a certain fraction of carriers in Ta_2O_5 form polaronic states at high temperature, which affects their conductivity and mobility. The discovery of the stability of polarons in this work provides fundamental insight into conduction processes in Ta_2O_5, which have not been completely understood yet.[34]

3.9 Conclusion

In conclusion, we use predictive atomistic calculations to investigate the thermodynamic, electronic, and optical properties of oxygen vacancies in Ta_2O_5. We found a correlation between the stability of oxygen vacancies at a certain site and the contribution of oxygen orbitals from that specific site to the charge density of the VBM state. Our calculations provide accurate and conclusive evidence that comprehensively explain the microscopic origin of a wide range of defect levels reported in various experiments and identify them with oxygen vacancy states. Defect states in the range of 0.6-0.8 eV reported in the TSC measurements and the *I-V* measurement with Poole-Frenkel conduction mechanism can be identified as thermodynamic transition levels of oxygen vacancies in Ta_2O_5 based on the calculated ionization energy of 0.81 eV. In addition, our configuration coordinate diagram indicates that defect levels are located around 1.56 eV below the CBM and 2.75 eV above the VBM, which describe consistently the defect states in photoconductivity, absorption, and photoluminescence measurements, once the exciton binding energy is considered. We show that oxygen vacancies in each oxidation states have totally different band structures and charge distributions, and the dominant oxidation state can be controlled with the growth conditions. We further show the formation of polarons and their complexes with oxygen vacancies, which can further the understanding on physical mechanisms of charge transports in Ta_2O_5. Our results provide fundamental understanding about the nature and characteristics of oxygen vacancies and can be applied to tune and optimize the electrical and optical properties of Ta_2O_5 films and the performance of devices.

3.10 References

1. Miao, F. et al. Anatomy of a Nanoscale Conduction Channel Reveals the Mechanism of a High-Performance Memristor. Advanced Materials 23, 5633–5640 (2011).

2. Park, G.-S. et al. In situ observation of filamentary conducting channels in an asymmetric Ta2O5−x/TaO2−x bilayer structure. Nature Communications 4, 1 (2013).

3. Lee, J., Du, C., Sun, K., Kioupakis, E. & Lu, W. D. Tuning Ionic Transport in Memristive Devices by Graphene with Engineered Nanopores. ACS Nano 10, 3571–3579 (2016).

4. Lee, J., Lu, W. D. & Kioupakis, E. Electronic and optical properties of oxygen vacancies in amorphous Ta2O5 from first principles. Nanoscale 9, 1120–1127 (2017).

5. Lau, W. S., Leong, L. L., Han, T. & Sandler, N. P. Detection of oxygen vacancy defect states in capacitors with ultrathin Ta2O5 films by zero-bias thermally stimulated current spectroscopy. Applied Physics Letters 83, 2835–2837 (2003).

6. Mikhaelashvili, V. et al. Electrical characteristics of metal-dielectric-metal and metal-dielectric-semiconductor structures based on electron beam evaporated Y2O3, Ta2O5 and Al2O3 thin film. Journal of Applied Physics 84, 6747–6752 (1998).

7. Fleming, R. M. et al. Defect dominated charge transport in amorphous Ta2O5 thin films. Journal of Applied Physics 88, 850–862 (2000).

8. Thomas, J. H., III. Defect photoconductivity of anodic Ta 2O 5films. Applied Physics Letters 22, 406–408 (1973).

9. Alers, G. B. et al. Nitrogen plasma annealing for low temperature Ta2O5 films. Applied Physics Letters 72, 1308–1310 (1998).

10. Markosyan, A. S., Route, R., Fejer, M. M., Patel, D. & Menoni, C. Study of spontaneous and induced absorption in amorphous Ta 2O 5and SiO 2dielectric thin films. Journal of Applied Physics 113, 133104 (2013).

11. Ivanov, M. V., Perevalov, T. V., Aliev, V. S., Gritsenko, V. A. & Kaichev, V. V. Electronic structure of δ-Ta 2O 5with oxygen vacancy: ab initiocalculations and comparison with experiment. Journal of Applied Physics 110, 024115–6 (2011).

12. Sawada, H. & Kawakami, K. Electronic structure of oxygen vacancy in Ta2O5. Journal of Applied Physics 86, 956–959 (1999).

13. Lee, J., Lu, W. & Kioupakis, E. Electronic properties of tantalum pentoxide polymorphs from first-principles calculations. Applied Physics Letters 105, 202108 (2014).

14. Xiao, B. & Watanabe, S. Oxygen vacancy effects on an amorphous-TaOx-based resistance switch: a first principles study. Nanoscale 6, 10169 (2014).

15. Zhu, L., Zhou, J., Guo, Z. & Sun, Z. Synergistic Resistive Switching Mechanism of Oxygen Vacancies and Metal Interstitials in Ta 2O 5. The Journal of Physical Chemistry C 120, 2456–2463 (2016).

16. Jiang, H. & Stewart, D. A. Enhanced oxygen vacancy diffusion in Ta 2O 5resistive memory devices due to infinitely adaptive crystal structure. Journal of Applied Physics 119, 134502–7 (2016).

17. Rinke, P. et al. First-Principles Optical Spectra for FCenters in MgO. Phys. Rev. Lett. 108, 126404–5 (2012).

18. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. Phys. Rev. Lett. 78, 1396–1396 (1997).

19. Blöchl, P. E. Projector augmented-wave method. Phys. Rev. B 50, 17953–17979 (1994).

20. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. Phys. Rev. B 59, 1758–1775 (1999).

21. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. Phys. Rev. B 47, 558–561 (1993).

22. Kresse, G. & Furthmüller, J. Efficient iterative schemes forab initiototal-energy calculations using a plane-wave basis set. Phys. Rev. B 54, 11169–11186 (1996).

23. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. J. Chem. Phys. 118, 8207–8215 (2003).

24. Van de Walle, C. G. First-principles calculations for defects and impurities: Applications to III-nitrides. Journal of Applied Physics 95, 3851–30 (2004).

25. Freysoldt, C., Neugebauer, J. & Van de Walle, C. G. Electrostatic interactions between charged defects in supercells. Phys. Status Solidi B 248, 1067–1076 (2010).

26. Freysoldt, C., Neugebauer, J. & Van de Walle, C. G. Fully Ab InitioFinite-Size Corrections for Charged-Defect Supercell Calculations. Phys. Rev. Lett. 102, 016402–4 (2009).

27. Janotti, A. & Van de Walle, C. G. Fundamentals of zinc oxide as a semiconductor. Rep. Prog. Phys. 72, 126501–30 (2009).

28. Freysoldt, C. et al. First-principles calculations for point defects in solids. Rev. Mod. Phys. 86, 253–305 (2014).

29. Rondinelli, J. M. & Kioupakis, E. Predicting and Designing Optical Properties of Inorganic Materials. Annu. Rev. Mater. Res. 45, 491–518 (2015).

30. Janotti, A., Franchini, C., Varley, J. B., Kresse, G. & Van de Walle, C. G. Dual behavior of excess electrons in rutile TiO 2. phys. stat. sol. (RRL) 7, 199–203 (2013).

31. Setvin, M. et al. Direct View at Excess Electrons in TiO2Rutile and Anatase. Phys. Rev. Lett. 113, 086402–5 (2014).

32. Bryksin, V. V., Goltsev, A. V., Khanin, S. D., Novotelnova, A. V. & Vasilev, A. N. Nonlinear Current—Voltage Characteristics of Ta2O5 and Nb2O5 Amorphous Oxides. Phys. Status Solidi B 161, 777–781 (1990).

33. Khanin, S. D. Kinetic Electronic Phenomena in Metal Oxide Dielectric Films. MSF 185-188, 563–572 (1995).

34. Chaneliere, C., Autran, J. L., Devine, R. A. B. & Balland, B. Tantalum pentoxide (Ta2O5) thin films for advanced dielectric applications. Materials Science and Engineering: R: Reports 22, 269–322 (1998).

Chapter 4 Charge Transition during Resistive Switching in Oxide-based RRAM devices

4.1 Introduction

It is challenging to directly observe the temporal CF growth in valence-change memory (VCM, commonly referred as oxide-RRAM), normally based on transition metal oxides, because the mobile species in VCM cells, oxygen vacancies (V_Os), are intrinsic defects and pose significant challenges for experimental analysis.[1,2] Even though it is generally accepted that resistive switching (RS) occurs due to the redistribution of charged oxygen vacancies driven by an external electric field, comprehensive atomistic models on internal dynamics of RS processes in oxide RRAM have been lacking. Specifically, there appear to be contradictions in the general description of the microscopic processes in RS. For example, oxygen vacancies involved in RS are generally considered as positively charged (V_O^{2+}) to explain the electric field-driven RS process (i.e., V_O^{2+} drift).[3] However, in this case, a strong repulsive electrostatic force between charged oxygen vacancies is expected, especially in filaments where the concentration of oxygen vacancy is very high. This in turn contradicts the experimentally observed long retention and stability of CFs in oxide-based RRAM devices. Developing an atomistic picture that can comprehensively explain the internal dynamics of RS in oxide-based RRAM thus becomes critical to improve the understanding of the RS processes and guide continued device optimization and application.

In this work, through predictive atomistic calculations in combination with electrical measurements under visible-light illumination, we attempt to understand the interactions between oxygen vacancies and the microscopic processes that are required for stable RS in Ta_2O_5-based

RRAM. We propose a model of a series of charge-transition processes during RS that explains the drift and aggregation of vacancies. The developed model was validated by experimental measurements in illuminated devices in which accelerated RS behaviors and reduced activation energy were observed due to light-assisted charge transition. Our results provide comprehensive understanding on the internal dynamics of resistive switching in oxide-RRAM and will help guide device optimization and applications.

4.2 Methodology

Computational methods: The equilibrium geometry of the atoms, the total energies, and electronic wave functions are obtained using the plane-wave pseudopotential method within the generalized gradient approximation (GGA) functional of Perdew, Burke, and Ernzerhof (PBE).[4] We used projector augmented-wave (PAW)[5,6] pseudopotentials as implemented in the Vienna ab-initio simulation package (VASP)[7,8] with a plane-wave cutoff of 400 eV and included the Ta $5s^2 5p^6 5d^3 6s^2$ and O $2s^2 2p^4$ electrons in the valence.

The formation energy of defects was calculated by the following equation[9]:

$$E^f[X^q] = E_{tot}[X^q] - E_{tot}[bulk] - \sum_i n_i \mu_i + q\mu_e + E_{corr} \quad (4.1)$$

where $E_{tot}[X^q]$ and $E_{tot}[bulk]$ are the total energy of a cell having an a defect with charge q and a defect-free bulk cell, respectively. The parameters μ_i and μ_e are the chemical potentials of atoms which are removed (μ_i<0) or added (μ_i>0) and the chemical potential of electron. The E_{corr} term includes the finite-supercell size correction for charged defects[10,11] and the electron and chemical reservoir corrections for the PBE-based formation energy[12]. A large supercell consisting of 252 atoms was employed to investigate a V_O dimer with long distance while minimizing artificial

interactions with periodic images in finite-size supercell approach. The binding energy was calculated with the PBE functional and a 1×1×3 supercell that allows the distance of a vacancy dimer to vary up to 15Å. The supercell used in this work was generated by duplicating an 84-atom cell that was verified in our previous work to exhibit consistent electronic and optical properties with various experimental data.[13,14]

Experimental Methods: Ta_2O_5-based memristive devices used in this work were fabricated in a crossbar structure on a SiO_2/Si substrate. The bottom electrode (Pd, 40nm) with an adhesion layer (NiCr, 5nm) was deposited by e-beam evaporation and pattered by photolithography and lift-off. Then, a tantalum suboxide (TaO_x) layer (40nm) was deposited by reactive sputtering using a Ta target with a gas flow of Ar/O_2 (32.3/1), followed by the deposition of tantalum pentoxide (Ta_2O_5) layer by radio frequency (RF) sputtering using a Ta_2O_5 ceramic target without breaking vacuum. Next, the top electrode (Pd, 10nm) was deposited and patterned as the bottom electrode. Finally, reactive ion etching (RIE) process was conducted to etch the oxide layers covering the bottom electrode and open the bottom electrode pads. A green-light laser (520 nm) was employed for the light-illumination experiment.

4.3 Interaction between Oxygen Vacancies

The binding energy of oxygen-vacancy pairs (i.e., the difference between the formation energy of a vacancy dimer and two isolated vacancies) at different distances, as indicated in Figure 4.1a, was calculated by the following equation[9,15]:

$$E_{\text{binding}}(q) = E^f\left[V_{O_1}^q + V_{O_2}^q\right] - \left(E^f\left[V_{O_1}^q\right] + E^f\left[V_{O_2}^q\right]\right) \tag{4.2}$$

where E^f is the formation energy of oxygen vacancies (Equation 4.1) and q is their charge state. The calculated binding energy as a function of distance between two vacancies is shown in Figure

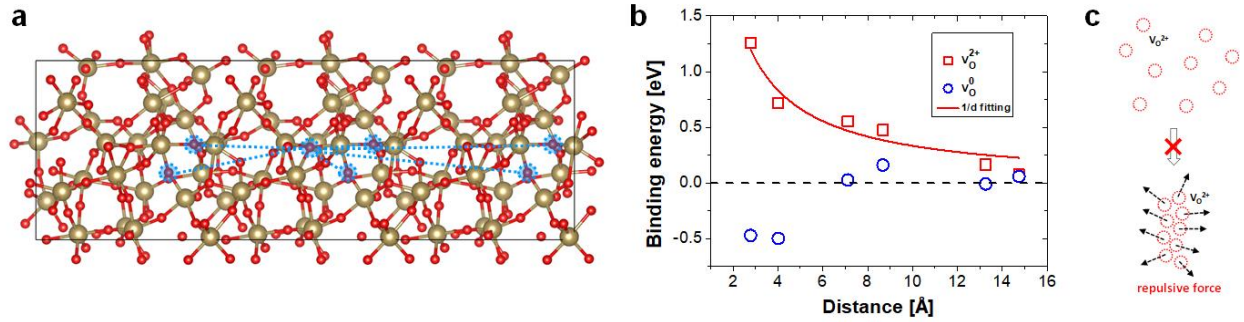

Figure 4.1: (a) Atomistic configuration of an amorphous-Ta_2O_5 supercell used to study the interactions of oxygen vacancy pairs with DFT calculations. Tantalum and oxygen atoms are represented by gold and red spheres, respectively. The investigated interactions between vacancy pairs at different distances are indicated by cyan circles and dashed lines. (b) Calculated binding energy of oxygen vacancy pairs as a function of distance. Charged vacancies show a strong Coulomb repulsion, while neutral vacancies exhibit a short-range attractive interaction. (c) A schematic illustration of the repulsive interaction between charged vacancies, preventing their aggregation for the formation of a CF in resistive switching, implying the instability of a CF consisting of charged vacancies.

4.1b. First, we found that the binding energy of vacancy pairs in the +2 charge state is positive, consistent with previous calculations in TiO_2,[15,16] and also increases with decreasing distance, indicating the expected strong repulsive interaction between charged vacancies. The repulsive energy is inversely proportional to the distance between the charged vacancies (V_O^{2+}s), consistent with electrostatic Coulombic repulsion between charged particles. It should be noted that although the repulsive interaction between charged vacancies becomes negligible at distances larger than ~13Å, the estimated distance between oxygen vacancies in CFs is typically much shorter. For example, J.Y. Chen *et. al* reported the O/Ta ratio of ~2:1 in the filament region in amorphous Ta_2O_5 based on the cross-sectional TEM and EDS analysis,[17] which corresponds to a high V_O concentration with a distance between vacancies of ~5Å. Therefore, a strong repulsive force appears between charged V_O^{2+}s for typical concentrations in CFs, preventing vacancies from aggregating and forming stable CFs as illustrated in Fig 1c. This contradicts the experimentally observed stable CFs in Ta_2O_5-based RRAM devices with long retention and high V_O concentration.[18,19]

On the other hand, our calculations show that neutral oxygen vacancies (V_O^0) exhibit a short-range attraction with a negative binding energy[15,16], as shown in Figure 4.1b. To understand the driving force of the attractive interaction between neutral vacancies, we determined the localization length of electrons in occupied midgap states induced by neural vacancies. Our results show that electrons bound to neutral vacancies are highly localized and their charge density decreases exponentially with increasing distance from the vacancy, as shown in Figure 4.2a, with a localization length of 4.8 ± 0.7Å. This localization length of neutral vacancies in amorphous

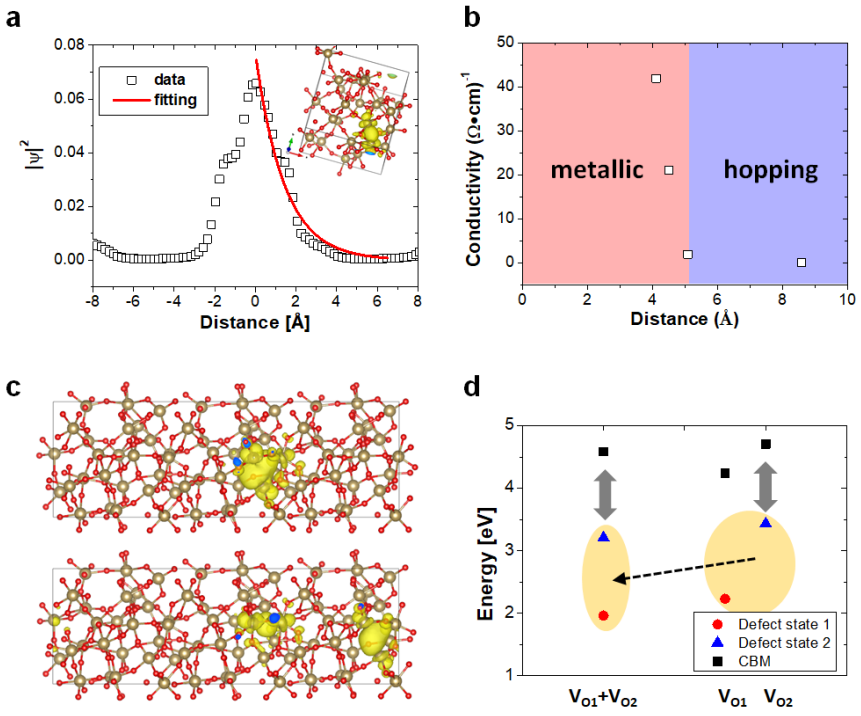

Figure 4.2: (d) Charge-density profile of a neutral oxygen vacancy as a function of the distance from the vacancy site, fitted with a decaying exponential function with a characteristic decay length of 4.8 ± 0.7Å. Inset: charge distribution of the V_O defect state. (e) Experimentally measured conductivity of tantalum oxide films of different stoichiometry (data from Ref. 20) as a function of distance between vacancies, showing different conduction mechanisms (metallic vs. hopping). The transition between the two mechanisms occurs at a vacancy distance of ~5Å, which is similar to the size of V_O wave function. (f) Charge distribution of oxygen vacancy pairs with different distances. Ttop: 2.7Å and bottom: 13.2Å. A significant overlap between the two V_O orbitals is observed for the short-distance pair, leading to hybridization. (g) Energy levels of defect states of an oxygen-vacancy dimer (V_{O1}+V_{O2}) and of two isolated vacancies, showing the lowering of two occupied defect states due to the hybridization and interaction with the conduction band. The energy levels are referenced to the valence band maximum.

Ta_2O_5 is remarkably consistent with the critical V_O distance for the transition between metallic and hopping conductions in Ta_2O_5 films and devices. Specifically, we plotted the experimentally measured conductivity of tantalum oxide films with different stoichiometry[20] as a function of V_O concentration and correspondingly the V_O distance (Figure 4.2b). In the regime of high V_O concentration ($>7\times10^{21}$ cm^{-3}), the conductivity of Ta_2O_5 films increases linearly with increasing V_O concentration while the trend of conductivity below the critical concentration exhibits not a linear but an exponential change, indicating metallic and hopping conduction mechanisms in these two regimes, respectively.[18,21,22] The corresponding critical distance (~5Å) is in very good agreement with the calculated localization length in our study, and also agrees with the experimentally estimated V_O distance in stable CF formation based on EDS analysis[17]. These agreements reveal the change of the conduction mechanism during RS in Ta_2O_5 devices. Specifically, metallic conduction of the on-state is caused by the continuous overlapping of localized electron states induced by V_O defects, where the V_O distance is shorter than the localization length, thus creating electronic conduction paths connecting one electrode to the other. V_O migration during the reset process reduces the V_O concentration and leads to increased V_O distance, resulting in reduced electron overlap. As a result, electrical conduction in the off-state is dominated by the hopping mechanism with a lower conductivity. The short-range attractive interaction between neutral vacancies (Figure 4.1b) can be explained by the hybridization of overlapping electron states.[23,24] The overlap of two electron states localized on different neutral oxygen vacancy sites (Figure 4.2c, top panel) gives rise to two mid-gap defect states that are lower in energy compared to those of isolated vacancies (Figure 4.2d) The lower-V_O level moves downward as a result of hybridization between the two defect orbitals,[24] while the upper-V_O level is also pushed down due to repulsion by the conduction band. Consequently, both occupied defect

levels become lowered in energy, leading to a negative binding energy and attractive interaction at vacancy-vacancy distances shorter than the localization length. This attractive interaction facilitates the aggregation of vacancies and the formation of stable CFs with high V_O concentration during RS.

4.4 Charge-Transition Model

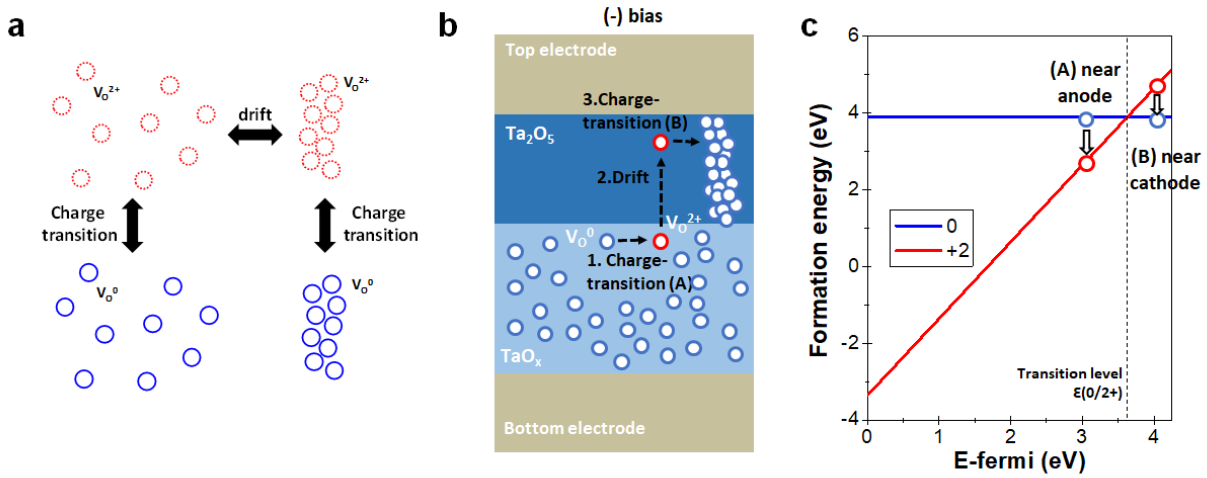

Figure 4.3: (a) Schematic of charge-transition and migration processes of oxygen vacancies in Ta_2O_5 memristive films. (b) Detailed illustration on the formation of a CF including the charge transition processes and ionic migration in a bilayer device. (c) Formation energy of an oxygen vacancy (neutral and +2 charged states) in amorphous Ta_2O_5 as a function of the Fermi level. Neutral vacancies near the anode transition to the charged state due to the lowered E_{Fermi} (A), followed by drift towards the cathode. The charged vacancies convert back to the neutral state as they approach the cathode because of the increased E_{Fermi} (B).

Even though neutral vacancies can explain the aggregation and the formation of stable CFs, their charge neutrality precludes interactions with an electric field and thus the field-driven drift as required during the RS. To resolve these contradictions, we consider a series of charge-transition processes during the RS process, which enables both the field-driven V_O drift and allows stable V_O aggregation. Figure 4.3a shows a schematic of the proposed charge transition and drift processes that result in CF formation/rupture during RS. Note that the formation energy of charged

defects (*e.g.*, V_O^{2+}) depends on the Fermi level, and that the formation energy of neutral vacancy dimers is within the range of the V_O^{2+} formation energy when the Fermi level changes from the valence band maximum (E_V) to the conduction band minimum (E_C). As a result, the stable V_O charge state can change depending on the Fermi level, which is in turn affected by the applied voltage during the device operation. For example, the lower Fermi level near the anode stabilizes the +2 charge state, while the neutral charge state becomes more stable near the cathode due to the elevated Fermi level. Figure 4.3b shows how isolated oxygen vacancies aggregate and form a CF during RS, along with the changes of relative stability between neutral and charged vacancies. Starting from the neutral charge state, isolated V_O^0s become ionized into V_O^{2+}s near the lower-Fermi-level anode. Then, the charged V_O^{2+}s migrate towards the cathode through the field-driven drift process and accumulate there. As more V_O^{2+}s accumulate near the cathode and the V_O concentration increases, the Fermi level rises, leading to additional charge transition from +2 to the neural state (Figure 4.3b). As a result, a stable CF consisting of aggregated V_O^0s without repulsive interactions is formed. If the applied field is removed without the subsequent charge transition process, the formed CF (consisting of V_O^{2+}s) becomes unstable. Therefore, this model can explain the experimentally observed excellent retention of Ta_2O_5-based devices,[19,25] providing an atomistic picture on the stability of CFs during RS and design guidelines to achieve desirable device properties such as long retention. For example, retention failure can occur when oxygen vacancies diffuse out from the CF by overcoming both the migration barrier and the binding energy with vacancies inside the CF in which vacancies attract each other. Indeed, the experimentally measured activation energy of 1.6eV through retention tests[19] is larger than the value of 1.2eV determined by profiling oxygen concentration during annealing in which diffusion is due to non-interacting vacancies.[26] This discrepancy can be explained by the increased barrier height due to

the attractive interaction energy. Therefore, resistive switching devices with long retention and low programming voltage can be achieved by utilizing defects with short-range attractive interactions.

4.5 Visible-Light Accelerated Resistive Switching Behaviors

To test this hypothesis of charge transition of oxygen vacancies during RS in oxide-based RRAM, we designed an experiment that employs visible-light illumination to control the V_O charge state. Specifically, electrons trapped at the mid-gap defect states of V_O^0 can be excited into the conduction band by light absorption, inducing a charge transition from 0 to +2 state. For example, it has been reported that V_O migration in $SrTiO_3$ is accelerated by visible-light illumination due to the photo-excitation of mid-gap states, where V_O^{1+}s in the form of Ti-V_O complexes are ionized to V_O^{2+}s with a lower activation energy for ionic migration.[27] In our experiment, we employed green light (520nm) whose photon energy is larger than the absorption energy of oxygen vacancies in Ta_2O_5 but smaller than the Ta_2O_5 band gap.[13,14] As a result, the characteristics of RS behaviors in Ta_2O_5-based RRAM devices would be affected by visible-light illumination if the charge-transition processes are indeed involved in the RS. Figure 4.4a shows the structure of a Ta_2O_5-based bilayer device used in this work, where a thin (10 nm) top electrode (TE) was used to allow light passthrough onto the switching layer. The 10 nm-Pd TE layer is indeed transparent to the green light used in this study, and devices with the thin TE are still fully functional and show reliable RS characteristics that are almost identical to standard devices with thick TE. A device being illuminated by a green-light laser (520 nm) during electrical measurements is shown in Figure 4.4a.

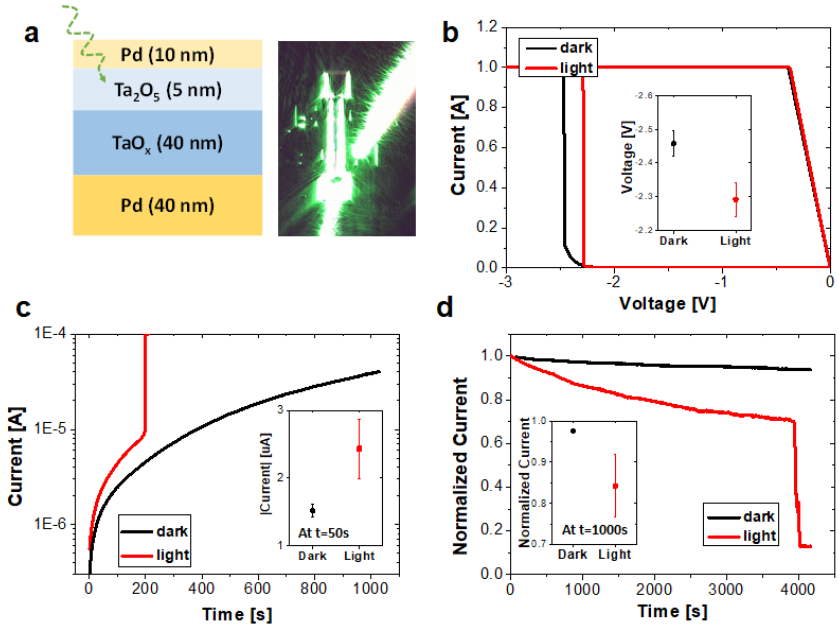

Figure 4.4: (a) A schematic of a Ta_2O_5-based bilayer device structure with thin top electrode to enable the light illumination on the switching layer (left). Optical-microscope image of the device illuminated by a green-light laser during electrical measurements (right). (b) DC *I-V* characteristics during electro-forming process with and without light illumination. Inset: average and standard deviation of forming voltage measured from multiple devices under light and dark conditions. (c) Current evolution of as-fabricated devices under constant voltage stress (-2V) with and without light illumination. Inset: current level of devices after 50s under the constant bias. (d) Current evolution of on-state devices under constant erasing voltage stress (0.8 V). Inset: current level after 1000s under the constant bias.

The effects of light illumination on the RS behavior were first examined in the electro-forming process. As shown in Figure 4.4b, illuminated devices show a smaller forming voltage at which the device conductance abruptly increases from the highly resistive virgin state. For example, the average forming voltage for the illuminated devices is -2.3V ± 0.05V, which is smaller than that of unilluminated devices, -2.45V ± 0.04V. To more clearly illustrate the light-induced effects, devices in virgin state were subjected to a low voltage (-2V) stress, where the applied voltage is lower than the typical forming voltage, and the current through the devices was monitored. Figure 4.4c shows the current evolution for devices with and without light illumination. The average current value for the illuminated device is clearly larger than the dark case, as further

verified in the inset of Figure 4.4c. In addition, an abrupt drop of resistance, corresponding to the electroforming process, occurs only under illumination conditions within our measurement time for such a relatively small bias.

Such behavior of the resistance decrease can be well understood in the picture of V_O-migration from the V_O-rich TaO_x layer to the V_O-poor Ta_2O_5 layer during the electroforming process. Since the Ta_2O_5 switching layer is much more resistive, the total device resistance is determined by the V_O concentration in the Ta_2O_5 layer. As charged oxygen vacancies migrate to the switching layer due to field-driven drift under the negative voltage applied to the top electrode, the total resistance of the device is reduced over time. Here, the light illumination speeds up the injection of charged V_O into the switching layer by facilitating the charge transition from V_O^0 to V_O^{2+}. That is, light absorption converts more neutral vacancies to the +2 charged state, with the V_O^{2+} responding to the applied electric field and migrate into the switching layer, leading to faster conductance increase and reduced electroforming voltage.

The effect of visible-light illumination was further investigated in the RESET process. Similarly, a constant voltage (0.8 V), which is lower than the typical RESET voltage, was applied to devices with and without light illumination. Prior to this measurement, the devices were switched to the LRS with similar resistance under dark conditions. Figure 4.4d shows the change of the normalized current over time under a constant positive 0.8 V voltage bias for the illuminated and dark cases. Both cases show the degradation of the LRS current over time. However, the illuminated devices show much faster decay, followed by an abrupt drop of current that corresponds to the breakage of the CF. The inset of Fig 3d shows the average value of current after 1000s under dark and illuminated conditions from multiple devices, verifying the accelerated erasing process under light illumination. The accelerated RESET transition by visible-light

illumination observed here is consistent with a previous work in which the electrical conductivity and the size of a nanoscale CF (formed by a conductive atomic force microscope (C-AFM) tip) in HfO_2 became reduced upon white-light illumination.[28]

The accelerated degradation of the LRS under illumination can also be explained by photo-induced charge transition from V_O^0 to V_O^{2+}. In general, the RESET process is believed to be driven by out-diffusion of oxygen vacancies due to Joule heating followed by V_O drift that create and widens a V_O-depleted gap.[3] However, the applied bias in this measurement itself is not large enough to cause neither diffusion nor drift, and thus no significant loss of V_Os from the CF. This is consistent with the slow degradation observed under dark conditions. On the contrary, light illumination leads to photo-excitation of trapped electrons at the mid-gap states of V_O^0s. As a result, the aggregated V_O^0s constituting the CF are transited to V_O^{2+}s that experience strong repulsive interactions, thus accelerating the migration of the vacancies and deteriorating the stability of the CF. Additionally, the light-assisted charge transition to the +2 state causes vacancies to respond to the applied electric field and enables V_O drift. Therefore, the observed faster decay under light illumination further supports our model of the dynamic RS involving the charge transition processes.

4.6 Transient Current Measurement

Additionally, the activation energies of ionic processes including ion migration and charge transition were experimentally determined through a transient current measurement.[29,30] Under constant positive bias, the as-fabricated devices exhibited transient current peaks as shown in Figure 4.5a. The positions of transient current peaks can be fitted with the Arrhenius equation, as shown in Figure 4.5b. We found that illuminated devices exhibit lower activation energy (0.42eV) than unilluminated devices (0.64eV), consistent with our experimental results showing light-induced accelerated RS behaviors (Figure 4.4). Specifically, since the charge transition from the

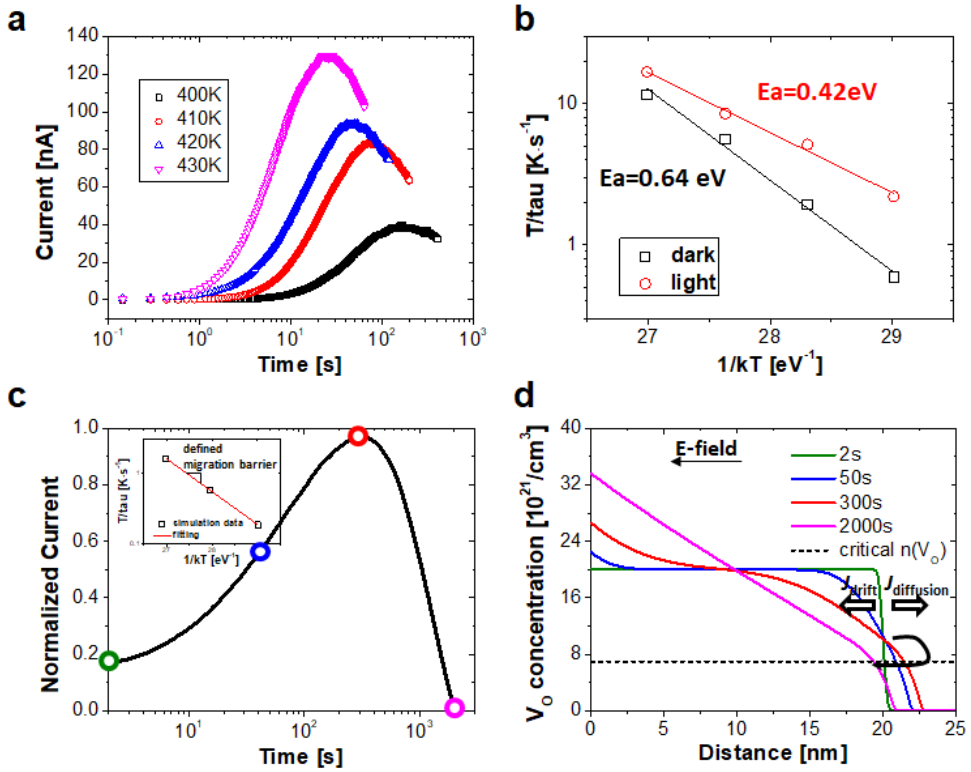

Figure 4.5: (a) *I-t* characteristics of as-fabricated devices under constant positive bias with light illumination at different temperature, showing transient current peaks (b) Arrhenius plot of the transient current peaks, showing lower activation energy under illumination. (c) Simulated I-t characteristics showing a transient current peak. Inset: Arrhenius plot of the current peaks simulated at different temperature. The activation energy determined from this plot is identical to the defined migration barrier in the numerical simulation setup. (d) Simulated V_O concentration profiles at each stage indicated in Figure 4.5c. The conductive region initially expands and subsequently contracts, leading to the current peak.

neutral to the +2 state is facilitated by light absorption, vacancies in an illuminated device are in the charged state and the estimated value from the transient current measurement is the activation energy for the migration of charged V_O^{2+}. Indeed, the experimentally estimated value of the activation energy under light illumination is in very good agreement with the theoretically calculated migration barrier of V_O^{2+} in Ta_2O_5 (~0.4eV)[31], supporting the charge-transition model. On the other hand, the activation energy measured under dark conditions corresponds to the charge-transition process, since this process dominates during the measurement of V_O^{2+} drift. We noted that the ionization energy of oxygen vacancy (0.6-0.8eV)[14,32,33] is indeed consistent with our experimental value.

An analytic model that reveals the relation between the peak position from the transient current measurements and the activation energy was developed as described in Equation (4.3):

$$\frac{T}{\tau} = \alpha D_0 \exp\left(-\frac{E_a}{kT}\right), \text{ where } \alpha = \frac{qE}{x_c k} \tag{4.3}$$

where T, D_0, E_a, x_c, k, and q are temperature, the preexponential constant and the activation energy of V_O diffusion, the farthest position (from the interface) that has the critical V_O concentration, Boltzmann's constant, and the charge of the oxygen vacancy, respectively. Here, the position of transient current peak (τ) was determined as the time at which the time derivative of the V_O concentration becomes zero, and thus diffusion and drift fluxes become equal with opposite sign, leading to either expansion or reduction of the conductive region. This analytic model was confirmed through numerical simulations in which the electronic continuity equation, the ionic drift/diffusion equation, and the Fourier equation for Joule heating are self-consistently solved.[3] With a simulation cell including V_O-rich and V_O-poor switching layers, a transient current peak was indeed observed under constant bias, as shown in Figure 4.5c. The activation energy

determined from the peak positions is identical with the set value for ionic migration barrier in this simulation (Figure 4.5c inset).

Figure 4.5d shows the concentration profile at each stage during the transient current measurement. In the beginning, diffusion is dominant due to the large V_O concentration gradient, leading to the motion of the boundary at which the V_O concentration is equal to the critical value towards the right, and thus to the expansion of the conductive region. The length of the conductive region is maximized when the diffusion and drift fluxes become equal, corresponding to the current peak (red circle in Figure 4.5c). Subsequently, with decreasing current the boundary starts to move towards the left, resulting in the transient current peak related to the activation energy of V_O migration as observed in the experiment and numerical simulation.

4.7 Conclusion

In summary, we show that positively charged oxygen vacancies (V_O^{2+}) in Ta_2O_5 experience strong mutual repulsion due to electrostatic Coulomb interaction, whereas neutral vacancies exhibit a short-range attractive force that facilitates aggregation to form a conductive filament. Neutral vacancies, however, do not drift in an applied electric field. We therefore believe a series of charge-transition processes need to occur to consistently explain both the aggregation and the field-driven drift of oxygen vacancies during RS. Our proposed model is supported by experiments under visible-light illumination. Specifically, memristor devices exhibit accelerated switching processes under illumination due to the light-assisted charge transition of oxygen vacancies, supporting the charge-transition model. Furthermore, we experimentally determined the activation energies of the charge transition and the V_O-migration processes based on the transient current measurement. Illuminated devices show a lower activation energy, with an experimentally

estimated value in good agreement with the calculated migration barrier of charge vacancies (V_O^{2+}). Additionally, we developed an analytic model verifying the relation between the transient current peak position and the activation energy, which is further supported by numerical simulations.

4.8 References

1. Yang, Y. & Lu, W. D. Progress in the Characterizations and Understanding of Conducting Filaments in Resistive Switching Devices. *IEEE Transactions on Nanotechnology* **15,** 465 (2016).

2. Lee, J. & Lu, W. D. On-Demand Reconfiguration of Nanomaterials: When Electronics Meets Ionics. *Advanced Materials* **30,** 1702770–33 (2017).

3. Kim, S., Choi, S. & Lu, W. Comprehensive physical model of dynamic resistive switching in an oxide memristor. *ACS Nano* **8,** 2369–2376 (2014).

4. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Phys. Rev. Lett.* **78,** 1396–1396 (1997).

5. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50,** 17953–17979 (1994).

6. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59,** 1758–1775 (1999).

7. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47,** 558–561 (1993).

8. Kresse, G. & Hafner, J. Ab initiomolecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B* **49,** 14251–14269 (1994).

9. Freysoldt, C. *et al.* First-principles calculations for point defects in solids. *Rev. Mod. Phys.* **86,** 253–305 (2014).

10. Freysoldt, C., Neugebauer, J. & Van de Walle, C. G. Fully Ab InitioFinite-Size Corrections for Charged-Defect Supercell Calculations. *Phys. Rev. Lett.* **102,** 016402–4 (2009).

11. Freysoldt, C., Neugebauer, J. & Van de Walle, C. G. Electrostatic interactions between charged defects in supercells. *Phys. Status Solidi B* **248,** 1067–1076 (2010).

12. Freysoldt, C. *et al.* Electron and chemical reservoir corrections for point-defect formation energies. *Phys. Rev. B* **93,** 165206–9 (2016).

13. Lee, J., Lu, W. & Kioupakis, E. Electronic properties of tantalum pentoxide polymorphs from first-principles calculations. *Applied Physics Letters* **105,** 202108 (2014).

14. Lee, J., Lu, W. D. & Kioupakis, E. Electronic and optical properties of oxygen vacancies in amorphous Ta2O5 from first principles. *Nanoscale* **9,** 1120–1127 (2017).

15. Kamiya, K. *et al.* ON-OFF switching mechanism of resistive–random–access–memories based on the formation and disruption of oxygen vacancy conducting channels. *Applied Physics Letters* **100,** 073502 (2012).

16. Kamiya, K. *et al.* Vacancy Cohesion-Isolation Phase Transition Upon Charge Injection and Removal in Binary Oxide-Based RRAM Filamentary-Type Switching. *IEEE Trans. Electron Devices* **60,** 3400–3406 (2013).

17. Chen, J.-Y., Huang, C.-W., Chiu, C.-H., Huang, Y.-T. & Wu, W.-W. Switching Kinetic of VCM-Based Memristor: Evolution and Positioning of Nanofilament. *Advanced Materials* **27,** 5028–5033 (2015).

18. Yang, J. J., Strukov, D. B. & Stewart, D. R. Memristive devices for computing. *Nature Nanotechnology* **8,** 13–24 (2013).

19. Choi, S., Lee, J., Kim, S. & Lu, W. D. Retention failure analysis of metal-oxide based resistive memory. *Applied Physics Letters* **105,** 113510–6 (2014).

20. Wei, Z. *et al.* Highly reliable TaOx ReRAM and direct evidence of redox reaction mechanism. in 1–4 (IEEE, 2008). doi:10.1109/iedm.2008.4796676

21. Goldfarb, I. *et al.* Electronic structure and transport measurements of amorphous transition-metal oxides: observation of Fermi glass behavior. *Appl. Phys. A* **107,** 1–11 (2012).

22. Zhang, Y. *et al.* Study of conduction and switching mechanisms in Al/AlO x/WO x/W resistive switching memory for multilevel applications. *Applied Physics Letters* **102,** 233502–5 (2013).

23. Carrasco, J., Lopez, N. & Illas, F. First principles analysis of the stability and diffusion of oxygen vacancies in metal oxides. *Phys. Rev. Lett.* (2004). doi:10.1103/PhysRevLett.93.225502

24. Kim, Y. S. & Park, C. H. Rich variety of defects in ZnO via an attractive interaction between O vacancies and Zn interstitials: origin of n-type doping. *Phys. Rev. Lett.* (2009). doi:10.1103/PhysRevLett.102.086403

25. Lee, M.-J. *et al.* A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O5−x/TaO2−x bilayer structures. *Nature Materials* **10,** 625–630 (2011).

26. Nakamura, R. *et al.* Diffusion of oxygen in amorphous Al 2O 3, Ta 2O 5, and Nb 2O 5. *Journal of Applied Physics* **116,** 033504–9 (2014).

27. Li, Y., Lei, Y., Shen, B. G. & Sun, J. R. Visible-light-accelerated oxygen vacancy migration in strontium titanate. *Scientific Reports* 1–7 (2015). doi:10.1038/srep14576

28. Zhou, Y. *et al.* White-light-induced disruption of nanoscale conducting filament in hafnia. *Applied Physics Letters* **107,** 072107–6 (2015).

29. Zafar, S. *et al.* Oxygen vacancy mobility determined from current measurements in thin Ba0.5Sr0.5TiO3 films. *Applied Physics Letters* **73,** 175–177 (1998).

30. Zafar, S., Jagannathan, H., Edge, L. F. & Gupta, D. Measurement of oxygen diffusion in nanometer scale HfO2 gate dielectric films. *Applied Physics Letters* **98,** 152903–4 (2011).

31. Jiang, H. & Stewart, D. A. Enhanced oxygen vacancy diffusion in Ta 2O 5resistive memory devices due to infinitely adaptive crystal structure. *Journal of Applied Physics* **119,** 134502–7 (2016).

32. Mikhaelashvili, V. *et al.* Electrical characteristics of metal-dielectric-metal and metal-dielectric-semiconductor structures based on electron beam evaporated Y2O3, Ta2O5 and Al2O3 thin film. *Journal of Applied Physics* **84,** 6747–6752 (1998).

33. Lau, W. S., Leong, L. L., Han, T. & Sandler, N. P. Detection of oxygen vacancy defect states in capacitors with ultrathin Ta2O5 films by zero-bias thermally stimulated current spectroscopy. *Applied Physics Letters* **83,** 2835–2837 (2003).

Chapter 5 Tuning Ionic Transport in Memristive Devices by Graphene with Engineered Nanopores

5.1 Introduction

As discussed in Chapters 1-4, the operations of memristors are based on internal atomic arrangements, e.g. migration and aggregation of oxygen vacancies (V_Os) that lead to the formation and rupture of filaments. In a standard memristor structure, however, the migration of oxygen vacancies is not well controlled.[1,2] For example, in a typical device an oxygen-deficient layer serving as the V_O supply and a near-stoichiometric, oxygen-rich layer serving as the switching layer are simply stacked on top of each other.[3,4] In this structure, oxygen ions (and correspondingly, V_Os) can migrate across the whole device area, leading to global injection of V_Os into the switching layer and causing several undesirable effects. First, the global V_O injection to the switching layer can lead to undesirably large leakage current in the switching layer and high operating currents. Second, the position and size of the CFs are random in each device and are determined mainly by film inhomogeneity, leading to large device-to-device and cycle-to-cycle variability issues. Finally, the over-injection of V_Os into the switching layer, if not controlled carefully (*e.g.*, through an external current compliance), can quickly lead to device failure in the form of stuck-at-1 (SA1) states since the over-supply of V_Os through the interface can lead to over-growth of the filament which becomes very difficult or impossible to reset.

In this study[5], we show that by inserting a graphene film (monolayer or multiple layer graphene) as an ion-blocking layer, ion migration that drives RS can be strongly modulated.

Furthermore, by selectively creating nanoscale openings with controlled sizes in the graphene ion-blocking layer, oxygen ions only selectively move through the engineered nanopores thus enabling systematic tuning of RS behaviors in a typical oxide-based memristor. Specifically, the operation current and RESET current levels are reduced, and more reliable device operation can be obtained in the graphene-inserted devices by preventing the over-injection of oxygen vacancies that results in SA1 issues. Extensive material analysis studies, including TEM and EDS, were employed to unambitiously verify the effects of graphene in resistive switching memory devices. The experimental findings were further supported by first-principles calculations of oxygen ion migration through nanoscale openings in graphene, confirming the role of graphene during RS at the atomic level. Finally, we show that reliable forming-free devices can be achieved in graphene-inserted devices after an annealing process, by promoting local ion diffusion only though engineered nanopores in the graphene layer that seeds controlled growth of filaments.

5.2 Methodology

Device fabrication. The Ta_2O_5-based memristive devices used in this work were fabricated in a crossbar structure on a SiO2 (100 nm)/Si substrate. The bottom Pd electrode (40 nm) with an adhesion layer of NiCr (5nm) was deposited and patterned with widths of 1-5 μm by e-beam evaporation (Cooke evaporator), photolithography (GCA AS200 Autostep), and lift-off processes. Next, the Ta_2O_5 layer (2-10 nm) was deposited by radio-frequency (RF) sputtering using a Ta_2O_5 ceramic target at room temperature. Monolayer graphene films grown on copper foil by chemical vapor deposition (CVD) were transferred onto the oxide layer using a standard PMMA-etchant method and patterned through photolithography and O_2 plasma etching. The top Ta/Pd electrode layers (40/20 nm) were deposited by direct current (DC) sputtering and e-beam evaporation,

respectively, followed by patterning process. Lastly, the bottom electrode pads were opened by etching the oxide layer through a reactive ion etching (RIE) process using SF_6/Ar.

Electrical measurements. A Keithely 4200 semiconductor characterization system and a custom-built electrical measurement system were used for the electrical measurements inside a probe station at ambient conditions. The bias voltage was applied to the TE with the BE being grounded.

TEM and EDS. Cross-sectional TEM samples were prepared by the standard lift-out process using a focus-ion beam system (FEI NOVA 200 NanoLab). TEM analysis was performed in the scanning mode (STEM) using a probe-corrected JEOL 2100F transmission electron microscope equipped with a EDS system.

Computational simulation. First-principles computational calculations were carried out based on density functional theory (DFT) using the Vienna *ab initio* simulation package (VASP). [30-32] The amorphous-Ta_2O_5 structures were generated by the melt and quench method as described in our previous work.[33] Atomic interaction and transport were simulated *via* first-principles molecular-dynamics (MD) simulations with the VASP code with a time step of 2 fs at a constant-temperature of 2500K.

5.3 Monolayer Graphene-inserted Memristor with Intrinsic pores

Graphene has recently been explored as an ion-blocking layer to control the ionic transport at nanoscale due to its extraordinary chemical and physical properties. Even though graphene is a single layer of carbon, it has been reported that monolayer graphene can completely block the penetration of ions and molecules due to its excellent impermeability.[6,7] In addition, graphene membranes with pores have shown promising results as size-selective barriers for water and gas purification, DNA sequencing, and single-molecule detectors where the desired particles can only

transport through pores in the graphene layer.[7-10] Graphene and amorphous carbon films have also been shown recently to be able to block ion migration in RS devices[11,12], although the effects were not systematically studied and the use of nanopores to modulate RS behaviors have not been experimentally attempted.

We examine how graphene films can be used to modulate RS switching behaviors and in particular how the RS effects can be controlled through engineered nanopores in graphene. Figure 5.1a shows the schematic of the memristive devices used in this work with a Ta/graphene/Ta_2O_5

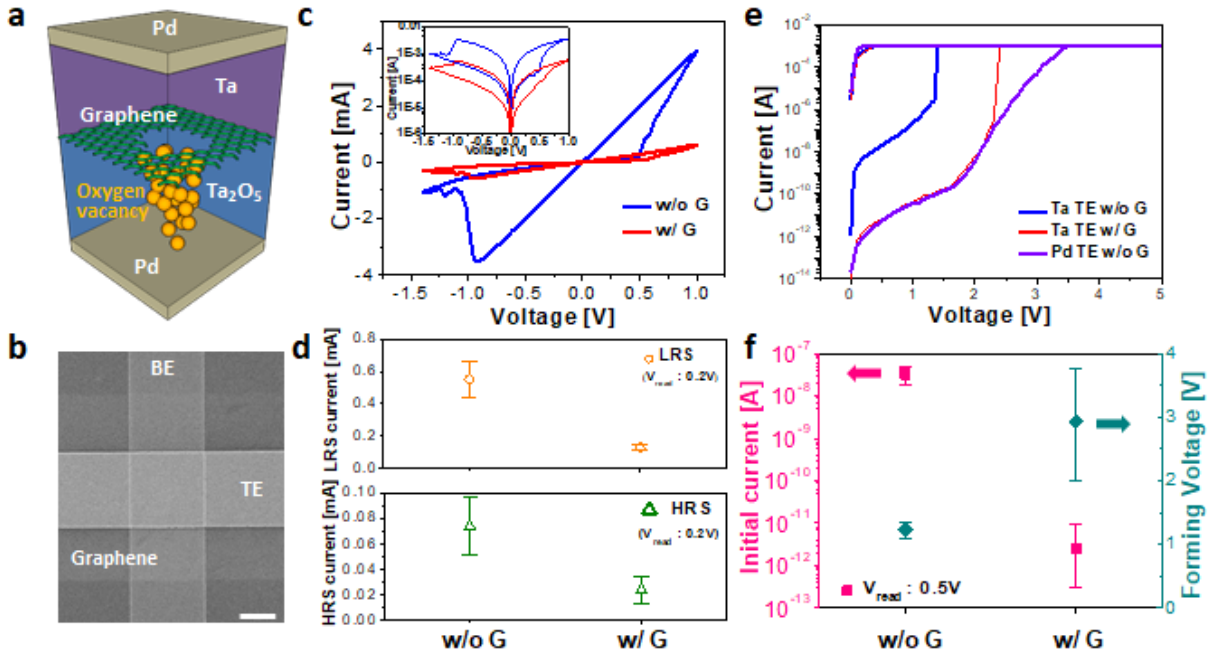

Figure 5.1: (a) Schematic of the graphene-inserted memristor structure where oxygen ions only transport through a nanopore created in the graphene layer, forming a CF with oxygen vacancies. (b) Top-view SEM image of the Ta/G/Ta_2O_5 device in a crossbar structure. Scale bar: 1 µm (c) Resistive switching *I-V* characteristic of the Ta/Ta_2O_5 and Ta/G/Ta_2O_5 devices (inset: log-scale). (d) Current levels in LRS (upper panel) and HRS (lower panel) of the Ta/Ta_2O_5 and Ta/G/Ta_2O_5 devices, showing the reduced operating current in the graphene-inserted devices and tighter variability control. The read voltage is 0.2 V. (e) *I-V* characteristic during the forming process of the as-fabricated Ta/Ta_2O_5 and Ta/G/Ta_2O_5 devices and a control device with an inert Pd top electrode. The resistance of the Ta/G/Ta_2O_5 device in the virgin state is similar to that of the Pd/Ta_2O_5 device, rather than the Ta/Ta_2O_5 device. (f) Distribution of the initial virgin sate current and the forming voltage for the Ta/Ta_2O_5 and Ta/G/Ta_2O_5 devices. The Ta/G/Ta_2O_5 devices are ~4 orders of magnitude more resistive than the Ta/Ta_2O_5 devices in the virgin state. The read voltage is 0.5 V. Reproduced with permission.[5] Copyright 2016, American Chemical Society.

($Ta/G/Ta_2O_5$) structure to study the effects of graphene on ionic transport and chemical reactions during resistive switching. In a conventional device consisting of a Ta/Ta_2O_5 structure, the thin Ta_2O_5 switching layer (5 nm used in this study) is sandwiched between an inert Pd bottom electrode (BE) and a reactive Ta top electrode (TE) where the Ta film also acts as an oxygen ion getting layer. It is widely known that spontaneous intermixing can occur at an interface between an oxide and a reactive metal electrode, such as Ti, Al, and Ta, due to the high reactivity of the metal layer. [4,13,14] This chemical interaction results in transport of oxygen ions (O^{2-}) from the oxide layer into the metal layer, and correspondingly the injection of V_Os into the oxide. Since the characteristics of resistive switching and the formation/rupture of conductive filaments are fundamentally driven by the generation, dissolution, and redistribution of V_Os in the switching layer, by inserting a graphene layer at the interface between the reactive Ta and the Ta_2O_5 switching layer in the $Ta/G/Ta_2O_5$ structure the ionic transport and redox reaction processes during resistive switching can be strongly affected, leading to selectively modified RS characteristics. Devices based on the $Ta/G/Ta_2O_5$ structure were fabricated in a crossbar structure, shown in Figure 5.1b, and extensively tested along with control devices based on the conventional Ta/Ta_2O_5 structure. Indeed, devices with even a monolayer graphene ($Ta/G/Ta_2O_5$ devices) exhibit significant and desirable changes in the *I-V* characteristics during RS compared with conventional Ta/Ta_2O_5 devices. First, the current levels of both the high resistance state (HRS) and the low resistance state (LRS) become lower in the (monolayer) graphene-inserted $Ta/G/Ta_2O_5$ devices, as shown in Figure 5.1c and 1d. While a typical Ta/Ta_2O_5 device requires ~4mA programming and RESET currents, the programming and RESET currents in a typical $Ta/G/Ta_2O_5$ device are reduced to ~0.6mA. Additionally, smaller standard deviations for the HRS/LRS currents in the $Ta/G/Ta_2O_5$ devices indicate the improved device-to-device uniformity (Figure 5.1d).

Additionally, more reliable device operation can be obtained in the graphene-inserted devices. For example, only a small percentage of Ta/Ta_2O_5 control devices (10-20%) showed repeatable RS behaviors without current compliance, while most devices became shorted and could not be erased either after the forming process or within the first a few cycles if a carefully designed current compliance was not applied. On the other hand, most Ta/G/Ta_2O_5 devices (80-90%) exhibited reliable switching without SA1 failure even without current compliance during the measurements, suggesting robust RS characteristics.

The effects of graphene can be more clearly observed in the initial state and during the forming process, as shown in Figures 5.1e-f. As seen in Figure 5.1e, the as-fabricated Ta/G/Ta_2O_5 device exhibits much lower conductance than the Ta/Ta_2O_5 device in the pre-forming, virgin state. In fact, the resistance of the Ta/G/Ta_2O_5 device in virgin state is almost identical to that of a Pd/Ta_2O_5 control device (purple line) and is 4 orders of magnitude higher than that of the Ta/Ta_2O_5 device. The forming voltage also increases from ~ 1V for the Ta/Ta_2O_5 to ~ 3V for the Ta/G/Ta_2O_5 device, as shown in Figure 5.1f. On the other hand, the Pd/Ta_2O_5 control device cannot be reliably formed and continued increase in forming voltage leads to hard breakdown effects instead of repeatable RS behaviors.

These different behaviors may be explained by the suppressed ion intermixing at the Ta/Ta_2O_5 interface due to the inserted graphene layer. In the Ta/Ta_2O_5 structure, V_Os generated in the Ta_2O_5 layer introduce defect levels within the forbidden gap of the oxide and contribute to increased leakage current in the film,[15] which can explain the higher initial conductance of the Ta/Ta_2O_5 device compared with the Pd/Ta_2O_5 device, where the lack of a reactive electrode prevents the formation of V_Os in the Ta_2O_5 layer. By inserting a graphene layer in the Ta/G/Ta_2O_5 device structure, the migration of oxygen ions into the Ta layer can be effectively blocked due to

the excellent impermeability of graphene, thereby suppressing the formation of oxygen vacancies in the switching layer through the interface. This hypothesis is consistent with the experimental results shown in Figure 5.1e, which shows that even with a reactive Ta electrode, the virgin-state resistance of the graphene-inserted device is similar to that of the device with an inert Pd electrode, rather than that of the Ta/Ta_2O_5 device. Additional tests on devices with different Ta_2O_5 layer

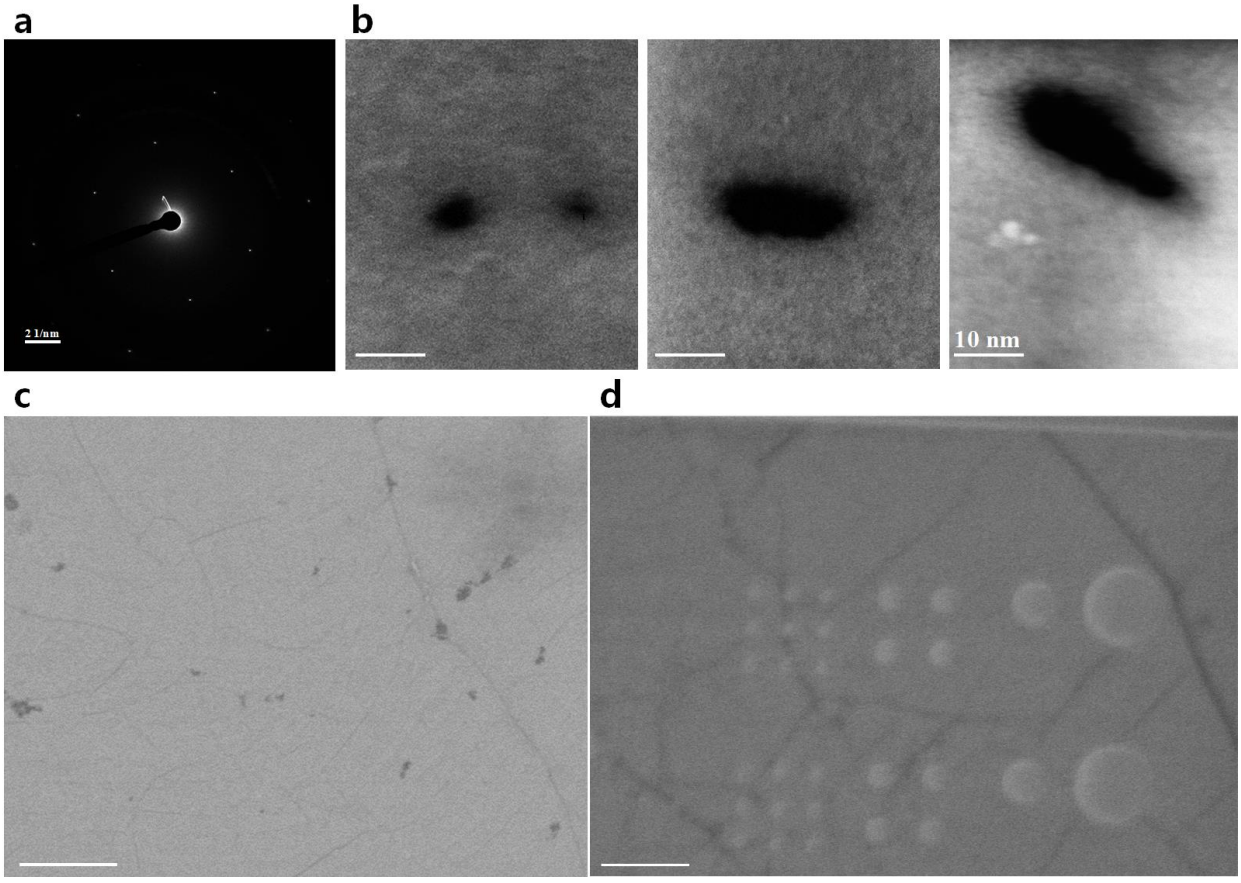

Figure 5.2: Characterization of the native and engineered nanopores in graphene layer. (a) Electron diffraction pattern of a transferred graphene film on TEM grid, where 6-fold symmetry of graphene is clearly visible. (b) TEM dark-field images showing several native nanopores in the transferred graphene film. Scale bar is 10 nm. (c) SEM image of a transferred graphene film on SiO_2/Si substrate. Scale bar is 2 µm. (d) SEM image of arrays of engineered nanopores created on the transferred graphene layer on the silicon oxide layer. Well-formed nanopores with designed size (50 – 400 nm) and locations can be clearly observed. The smallest nanopore (25nm) could not be observed by SEM due to the limit of resolution. Scar bar is 500 nm. Reproduced with permission.[5] Copyright 2016, American Chemical Society.

thicknesses further prove that the current through these devices in the virgin state is dominated by the bulk resistance of the Ta_2O_5 film, instead of interfacial effects.

The Ta/G/Ta_2O_5 devices show reduced current levels during RS, which may be explained by the formation of smaller or fewer CFs compared to the Ta/Ta_2O_5 devices. During SET, a positive voltage is applied on the Ta TE that attracts oxygen ions from and moves oxygen vacancies into the Ta_2O_5 layer. The accumulation of V_Os lead to the formation of V_O-rich CFs and the eventual switching of the device from the HRS to the LRS.[16] The reverse process occurs during RESET. When a graphene layer is inserted between the Ta TE and the Ta_2O_5 switching layer, oxygen ionic transport is only possible through holes (openings) in the graphene layer. It has been reported that native defects in the form of nanopores exist in CVD-grown graphene with pore size of ~10 nm and density of ~1 pore/μm[3,17] The existence of such native pores was indeed confirmed in the graphene film used in this study (Figure 5.2b). With a device size of 2 × 2 μm^2, we expect to have only a few nanopores on average in a single device, where oxygen ion migration between the Ta and Ta_2O_5 layers only takes place through the nanopores. As a result, the CFs can only grow from the nanopore regions and the size and position of the CFs will be determined by the size and position of the nanopores. Additionally, this local seeding effect further prevents over-injection of V_Os into the switching layer and can explain the improved reliability in the Ta/G/Ta_2O_5 devices. In the conventional structure without the graphene ion-blocking layer, V_Os can be injected into the Ta_2O_5 switching layer across the whole Ta/Ta_2O_5 interface and flood the switching layer. As a result, very thick CFs can be formed due to the injection of excess V_Os, making it difficult or impossible to RESET the device and causing the device to be stuck at the LRS unless a careful feedback mechanism such as current compliance during programming is implemented. By limiting the ion injection only at the localized openings through the graphene layer, the over injection

problem can be effectively mitigated resulting in more reliable device operations in the Ta/G/Ta_2O_5 devices.

5.4 Multilayer Graphene-inserted Memristor with Engineered Nanopores

To further test our hypothesis on the role of graphene as ion-blocking layer and the effects of local-seeding through nanopores, we fabricated Ta/G/Ta_2O_5 devices with multilayer graphene (MLG) stacks through a multiple graphene-transfer process. By transferring the graphene layer multiple times (*e.g.*, 3 times used in this study), we expect the native nanopores in the graphene layer to be completely blocked since the nanopores occur randomly and there is no position registry between nanopores in different layers during the transfer process. As a result, ion intermixing and ionic transport should be blocked completely in the Ta/MLG/Ta_2O_5 devices. Indeed, such devices cannot be successfully formed and RS behavior cannot be observed even at bias voltages up to 5 V and further increasing the forming voltage only leads to hard breakdown effects and SA1 failures. Ta/MLG/Ta_2O_5 devices with controlled artificial nanopores were then fabricated by creating holes in the transferred MLG layer *via* e-beam lithography and oxygen plasma etching, where the size of the openings was controlled to be 25 nm – 400 nm in diameter in different devices (Figure 5.2c). Significantly, with this approach, we can selectively create 1 artificial nanopore in one device, with controlled nanopore size. Figure 5.3a shows the initial current of the Ta/MLG/Ta_2O_5 devices with different nanopore sizes (with 1 nanopore per device) and control samples without engineered nanopores and Pd/Ta_2O_5 devices. In general, the MLG devices show an increase in current in the virgin state as the size of the nanopore is increased, consistent with our hypothesis that oxygen ion transport and intermixing only occur through the engineered nanopore in the device, where the pre-forming conductance is dependent on the intermixing area.

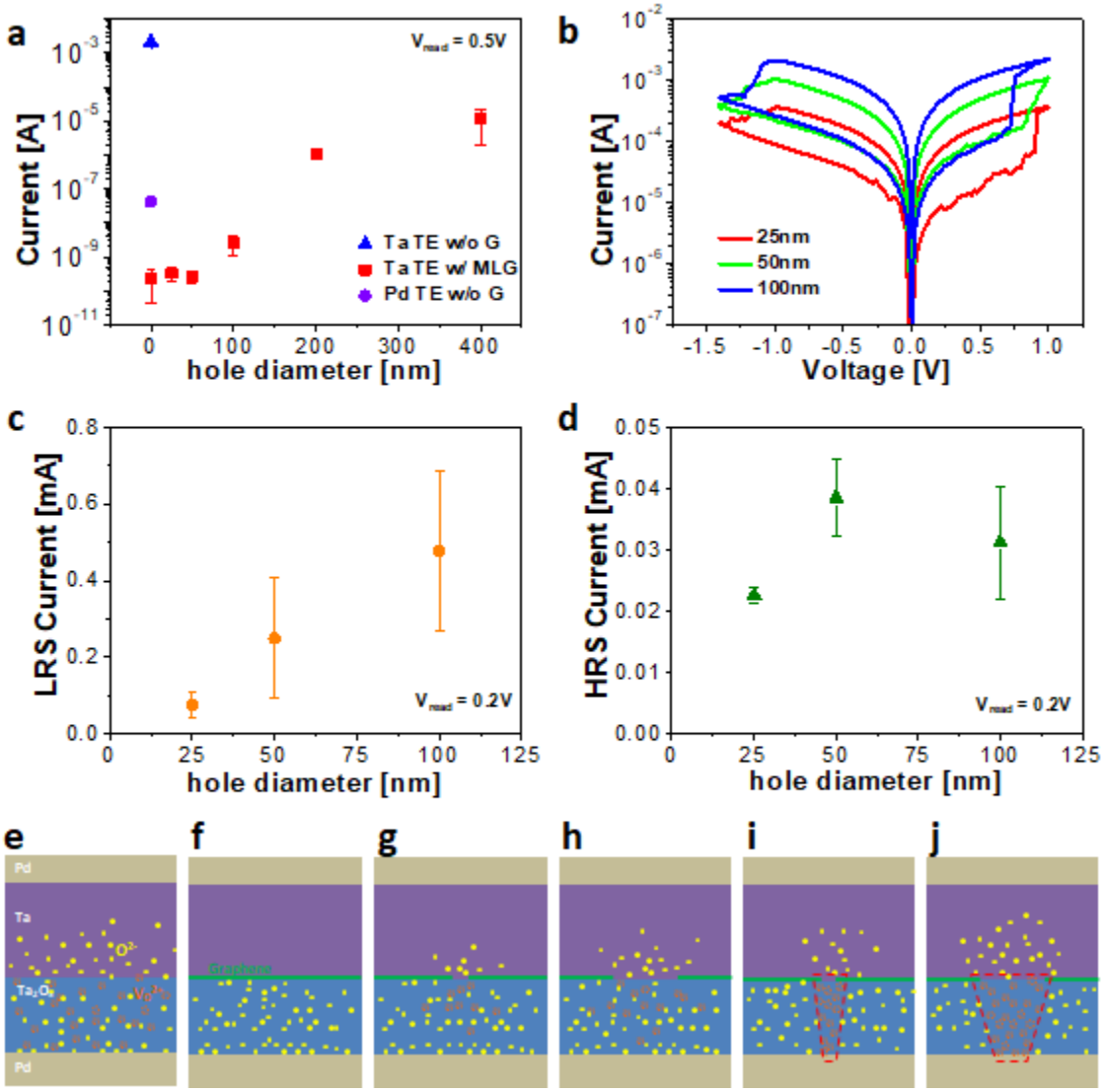

Figure 5.3: (a) Virgin state currents of the Ta/MLG/Ta_2O_5 devices having different sized nanopores (holes) created in the graphene layer, along with results from control samples having Ta and Pd top electrodes without the graphene layer. (b) Resistive switching *I-V* characteristic of the Ta/MLG/Ta_2O_5 devices with different sized nanopores fabricated in the graphene layer, showing that the switching characteristics can be systematically tuned by graphene with nanopores. (c) Current levels in LRS and (d) HRS of the graphene-inserted devices with different nanopore sizes. Read voltage is 0.2 V. (e-h) Schematic illustration of the spontaneous intermixing between the Ta and Ta_2O_5 layers in the in the Ta/Ta_2O_5 (e), Ta/MLG/Ta_2O_5 (f), and Ta/MLG/Ta_2O_5 with small/large nanopore (g,h), showing the fully consumed oxide by the intermixing, the complete blocking by the MLG, and the localized ionic transport induced by nanopore in the MLG, respectively. (i, j) Schematic illustration of the formation of CF in the MLG-inserted devices with nanopores where the size of CF is limited by the nanopore. Reproduced with permission.[5] Copyright 2016, American Chemical Society.

This trend is not obvious for devices with the smallest nanopores (25 nm and 50 nm diameters) which show similar virgin state conductance as the control samples without engineered nanopores.

This effect can be explained by the fact that the opening area in these devices is too small to allow extensive intermixing at virgin state, and the resistance of the device is dominated by the bulk resistance of the Ta_2O_5 switching layer, as in the case of the device without engineered nanopores. More importantly, systematically different RS behaviors are observed in the Ta/MLG/Ta_2O_5 devices with different artificial nanopore sizes, as shown in Figure 5.3b, with smaller pore sizes leading to lower programming current and higher LRS resistance. Figure 5.3c shows the LRS conductance (measured at a read voltage of 0.2 V) vs. the engineered nanopore size, and an approximately linear relationship can be obtained.

This effect again can be explained by the locally seeded growth of CF in graphene-inserted devices, as schematically shown in Figure 5.3e-j. Without the ion-blocking layer, oxygen ions can diffuse across the Ta/Ta_2O_5 interface throughout the device area and lead to high leakage current at the virgin state due to over-injection and SA1 failure (Figure 5.3e). By inserting MLG which completely blocks ion migration, V_Os cannot be generated in the Ta_2O_5 layer and the device cannot be formed and switched (Figure 5.3f). By creating an artificial nanopore in the MLG ion blocking layer, V_Os can be injected only locally through the nanopore, leading to more controlled V_O migration and suppressed leakage current in the virgin state (Figure 5.3g) and controlled filament growth during SET (Figure 5.3i). By increasing the nanopore size, both the leakage current in the virgin state (Figure 5.3h) and the filament size (Figure 5.3j) are increased due to increased oxygen ion diffusion through the increased opening at the Ta/Ta_2O_5 interface. The trend is not as clear in HRS (Figure 5.3d) which is more strongly affected by the bulk properties of the switching layer, although the devices with the smallest engineered nanopore size (25 nm) still showed the highest

HRS resistance. In addition, the MLG-inserted devices with artificial nanopores showed much smaller variation in the forming voltage and the virgin-state currents (Figure 5.3a) compared to the Ta/G/Ta_2O_5 devices using native nanopores (Figure 5.1f), indicating the improved controllability in engineered nanopore devices since the number of nanopores and pore size are randomly distributed in the native nanopore-based devices. Compared with standard Ta/Ta_2O_5 devices, the MLG-inserted device with the engineered nanopore showed better switching uniformity with lower operating currents for both the HRS and the LRS, again consistent with our explanation that the inserted-graphene with nanopores can reduce operating currents by confining the filament size with local injection of V_Os and also improve cycle-to-cycle uniformity by controlling the location of the filament. (Figure 5.4).

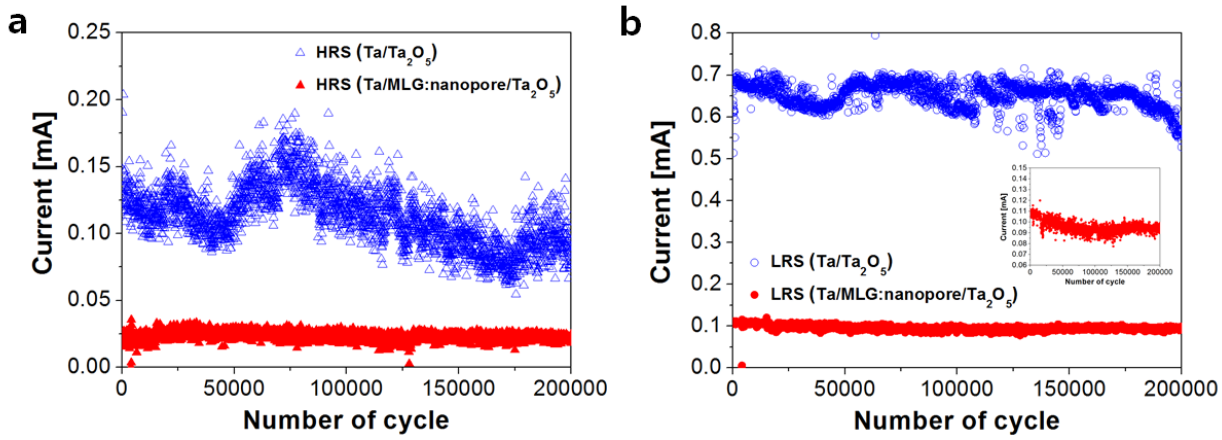

Figure 5.4: Endurance test of the Ta/Ta_2O_5 and Ta/MLG:nanopore/Ta_2O_5 devices for the HRS (a) and the LRS (b). (inset: the LRS current in enlarged scale) Reproduced with permission.[5] Copyright 2016, American Chemical Society.

The LRS/HRS current levels of the MLG-inserted device with an engineered 25 nm nanopore are similar to those of the monolayer graphene-inserted devices without artificial pores, indicating that the total area of the intrinsic nanopores in the monolayer graphene film is comparable to that of a 25 nm diameter pore, which is consistent with the estimated native pore size discussed earlier[17]. The non-uniformity induced by intrinsic nanopores in the monolayer

graphene case was examined by measuring the properties of devices with different electrode sizes, and the density of native nanopores was roughly estimated to be 1 nanopore per a few square micrometers. When the size of the artificial nanopore becomes too large, such as in the 200 and 400 nm diameter cases, the devices easily become stuck at the LRS and cannot be RESET, similar to results from the Ta/Ta_2O_5 control samples, as the large openings at the Ta/Ta_2O_5 interface are no longer efficient in preventing the over-injection of oxygen ions and the devices behave similarly to the conventional structure. We believe future studies should be based on engineered nanopores with even small pore sizes (*e.g.* < 10 nm) using more advanced fabrication techniques[18] to allow continued optimization in aggressively scaled devices. Additionally, continued optimization of the switching layer (e.g. the Ta_2O_5) layer itself remains an important topic. By controlling both the ion injection process using techniques discussed here and the filament growth process through switching layer optimizations, more uniform switching with desired switching characteristics including power, on/off ration can be obtained.

5.5 Microscopic and Spectroscopic Analysis on the Role of the Graphene Layer

To directly verify the effects of graphene on ionic transport at the Ta/Ta_2O_5 interface, transmission electron microscopy (TEM) and energy-dispersive X-ray spectroscopy (EDS) studies were carried out. To exclude the effects of intrinsic holes and cracks and to make the layers more visible in the cross-sectional elemental analysis, monolayer graphene was transferred several times (typically ~3) to achieve a multilayer graphene (MLG) stack that can completely block the oxygen ionic transport and intermixing at the Ta/Ta_2O_5 interface. A cross-section of the sample was then obtained for TEM analysis. The MLG layer in the $Ta/MLG/Ta_2O_5$ sample can be clearly identified between the Ta and Ta_2O_5 layers in the scanning TEM (STEM) annular dark-field (ADF) and

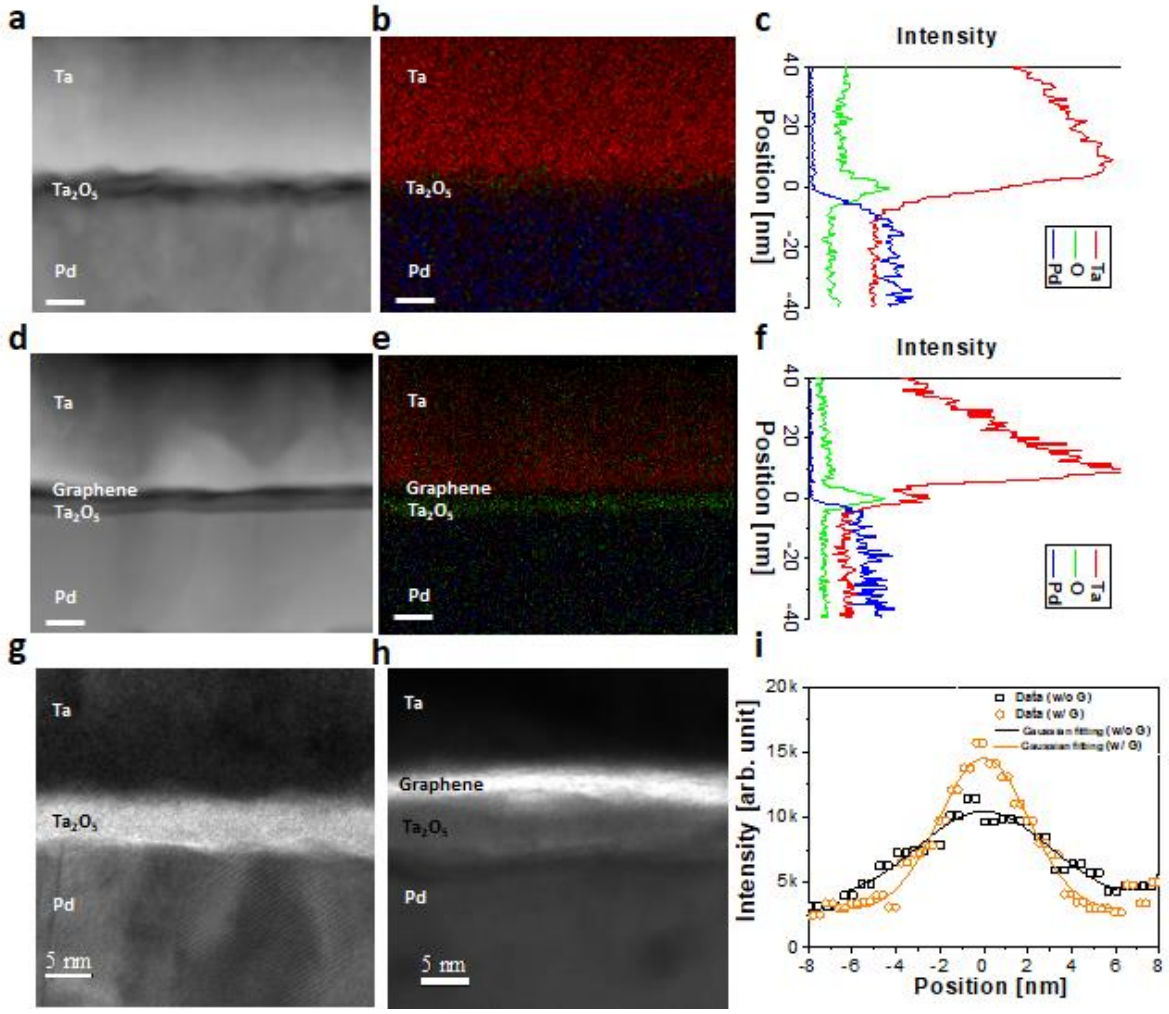

Figure 5.5: (a-f) Cross-sectional STEM dark-field images (a, d) and EDS elemental mapping of Ta (red), O (green), and Pd (blue) elements (b, e) in the Ta/Ta_2O_5 and Ta/MLG/Ta_2O_5 structure (scale bar: 10 nm), respectively, where the oxide layer is clearly separated from the reactive Ta electrode only in the graphene-inserted structure. Intensity profiles of Ta, O, and Pd elements along the vertical direction in the Ta/Ta_2O_5 (c) and Ta/MLG/Ta_2O_5 (f) structures. A Ta signal peak is not discernable in the oxide layer region for the Ta/Ta_2O_5 structure, indicating the intermixing between the Ta and Ta_2O_5 layers. (g, h) High-resolution STEM bright-field images of the Ta/Ta_2O_5 (g) and Ta/MLG/Ta_2O_5 (h) devices. (i) Oxygen EDS profiles along the vertical direction for the Ta/Ta_2O_5 and Ta/MLG/Ta_2O_5 devices. A narrower oxygen distribution is observed in the Ta/G/Ta_2O_5 case. Reproduced with permission.[5] Copyright 2016, American Chemical Society.

bright-field (BF) images (Figure 5.5d and h, respectively) compared to the Ta/Ta_2O_5 samples (Figure 5.5a and g, respectively). Figures 5.5b-c and e-f show the elemental mapping of Ta, Pd, and O elements and their intensity profiles from EDS measurements for the Ta/Ta_2O_5 and

Ta/MLG/Ta_2O_5 samples, respectively, where the profiles were obtained by summing up signals of each elements along the direction parallel to interface. In the Ta/Ta_2O_5 sample, the thin Ta_2O_5 layer cannot be clearly distinguished from the Ta electrode in the elemental mapping image, and the profile of Ta increases continuously from the substrate to the Ta layer without forming a peak for the oxide layer (Figure 5.5c), suggesting the spontaneous intermixing between the Ta and Ta_2O_5 layers. However, in the MLG-inserted sample the Ta_2O_5 layer can be distinctly separated from the Ta electrode in the elemental mapping image, as shown in Figure 5.5e. There is an apparent gap, consisting of C atoms and corresponding to the graphene layer, between the Ta and Ta_2O_5 layers , and a Ta concentration peak can be observed in Figure 5.5f corresponding to a well-defined Ta_2O_5 oxide layer, in contrast to the Ta/Ta_2O_5 case (Figure 5.5c). Figure 5.5i shows the profile of the oxygen signal in the Ta_2O_5 layer for both samples. The oxygen profile can be fitted with a Gaussian curve in both cases, where the Ta/Ta_2O_5 sample shows a much wider distribution of oxygen with a full width half maximum (FWHM) of 7.1 nm compared to the Ta/MLG/Ta_2O_5 case with a FWHM of 4.8 nm. Overall the EDS data unambiguously support our hypothesis that graphene can block the oxygen ion migration and intermixing at the interface, and thereby, enabling the modulation of oxygen ion transport from the oxide layer into the metal layer.

5.6 Atomistic Simulations on Ionic Transport

Atomistic-level simulations based on first-principles molecular dynamics (MD) were further performed to verify that ion transport occurs only locally through nanopores in the graphene layer. Figure 5.6a shows the atomic configurations of the three cells used in this simulation: Ta/Ta_2O_5, Ta/G/Ta_2O_5, and Ta/G:nanopore/Ta_2O_5 where crystal bcc-Ta and amorphous-Ta_2O_5 structures were used (see Methods). To speed up the ion diffusion which

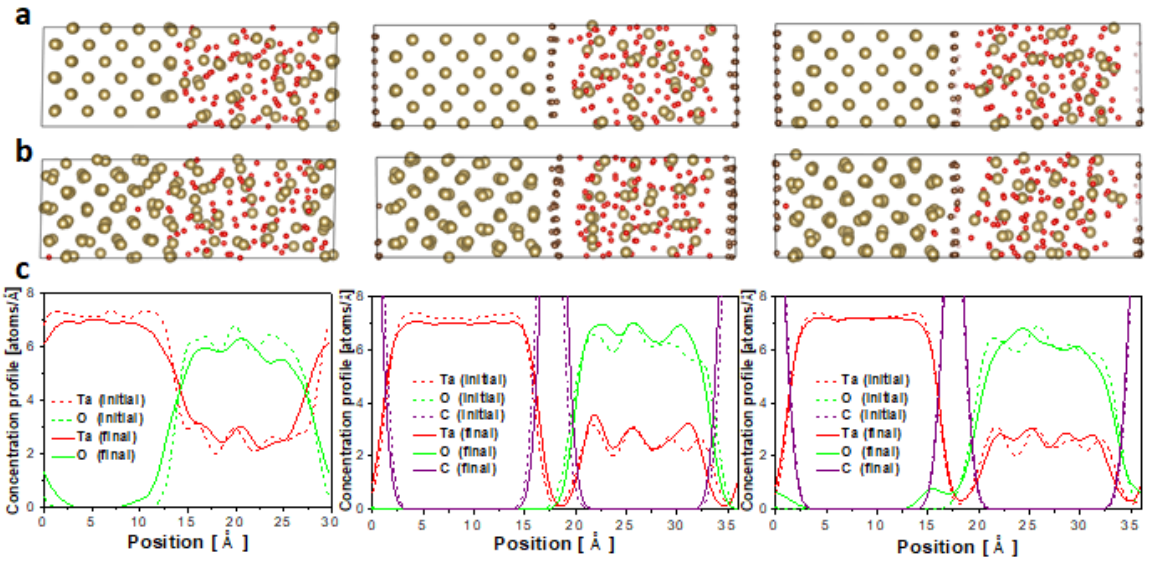

Figure 5.6: (a) Atomistic configurations of the three cells used in first-principles MD simulations: Ta/Ta_2O_5 (left), Ta/G/Ta_2O_5 (center), and Ta/G:nanopore/Ta_2O_5 (right) where crystal bcc-Ta and amorphous-Ta_2O_5 structures are employed. Ta, O, C, and H atoms are colored in gold, red, brown, and soft purple, respectively. (b) Final atomic configurations after 12 ps at 2500 K. In the Ta/Ta_2O_5 cell (left) there is clear intermixing at the interface between the Ta and Ta_2O_5 layers, causing the distorted arrangement of the crystalline Ta layer. In the Ta/G/Ta_2O_5 cell (center) the two layers are still distinctly separated by the inserted-graphene without any observed intermixing or chemical reactions due to graphene's excellent impermeability and chemical stability. In the Ta/G:nanopore/Ta_2O_5 case (right), transport of oxygen ions is limited and only takes place through the nanopore. (c) Atomic concentrations of Ta, O and C in the three cells before and after annealing. A significant increase in oxygen concentration in the Ta layer can be observed in the Ta/Ta_2O_5 case after annealing, while a smaller increase is observed in the Ta/G:nanopore/Ta_2O_5 case and no increase is observed in the Ta/G/Ta_2O_5 case. Reproduced with permission.[5] Copyright 2016, American Chemical Society.

exponentially depends on temperature according to the Arrhenius relation and observe noticeable ion migrations within the limited simulation time of first-principles MD, the trajectories of ions were simulated at a high temperature below the melting point of Ta for 12 picoseconds and the final atomic configurations are shown in figure 5.6b. Figure 5.6c shows the concentration profiles of Ta, O, and C atoms along the direction perpendicular to the film interfaces for the initial and final states for the three cells. Noticeable diffusion of oxygen ions from the oxide to the Ta metal layer is observed at the Ta-side of the interface in the Ta/Ta_2O_5 cell. In addition, the intermixing due to the exchange of Ta and O ions between the two layers can be seen over the whole device

area at the interface, consistent with our TEM and EDS results and also a recent report on the possibility of Ta cation motion.[23] In contrast, no intermixing and no ion transport across the interface were observed in the Ta/G/Ta_2O_5 cell where the metal and the oxide layers are clearly separated by graphene, verifying that oxygen ions cannot pass through the defect-free graphene film within our simulation time. When nanopores (openings) are created in the graphene layer, transport of oxygen ions into the metal layer can be observed again, confirming our local seeding hypothesis. These simulation results indicate that ion-transport can be controlled by nanopores in graphene, thus limiting the amount of oxygen vacancies injected in the resistive switching layer and modifying the RS behaviors in oxide memristive devices.

5.7 Electroforming-free Behavior

Finally, the effects of graphene on ion migration were further examined experimentally by annealing samples at a high temperature (300°C). Figure 5.7a shows the virgin state device resistance before and after annealing for different annealing times, for both Ta/Ta_2O_5 devices and Ta/G/Ta_2O_5 devices where a monolayer graphene is used in the Ta/G/Ta_2O_5 devices. Both types of devices show a decrease of the virgin state resistance with annealing time, indicating the diffusion of oxygen ions into the Ta layer during annealing and the injection of V_Os into the Ta_2O_5 layer causing increased leakage current in the Ta_2O_5 layer. However, the resistance of the Ta/Ta_2O_5 devices decreases drastically and becomes very conductive with a virgin state resistance of only a few hundred Ohms after just a short annealing time; while the virgin state resistance for the graphene-inserted devices decreases much slowly during annealing. Such behaviors are consistent with our explanations on the role of graphene as the ion-blocking layer and the locally-allowed diffusion paths (in this case native nanopores) through the graphene layer. Unlike the Ta/Ta_2O_5

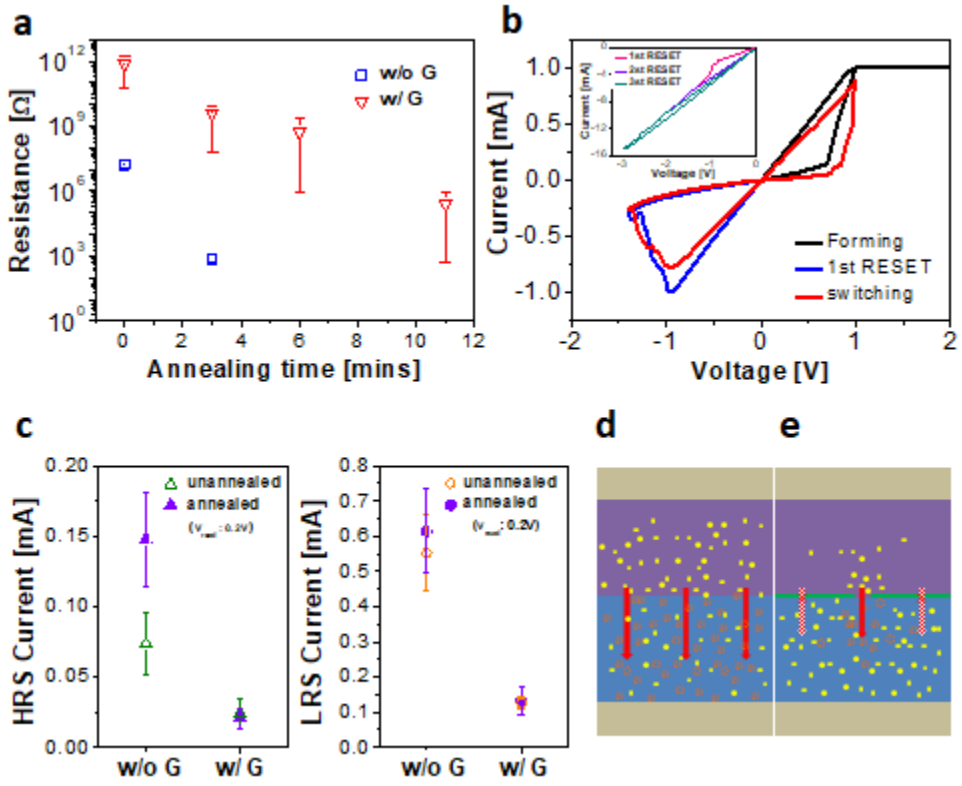

Figure 5.7: (a) Virgin-state resistance as a function of annealing time for devices with and without graphene. (b) Resistive switching *I-V* characteristic of the annealed Ta/G/Ta_2O_5 device, showing forming-free behavior. (inset: reset failure of the conventional Ta/Ta_2O_5 device after annealing.) (c) HRS (left) and LRS (right) current distribution before and after annealing for the Ta/Ta_2O_5 and Ta/G/Ta_2O_5 devices. No significant changes were observed in the Ta/G/Ta_2O_5 devices compared, while the HRS current is significantly increased for the Ta/Ta_2O_5 devices after annealing. (d, e) Schematic illustrations of the oxygen ion diffusion and the accompanying V_O injection processes for the Ta/Ta_2O_5 and Ta/G/Ta_2O_5 devices during annealing, respectively. The controlled and localized V_O injection in the Ta/G/Ta_2O_5 leads to reliable forming-free behavior. Reproduced with permission.[5] Copyright 2016, American Chemical Society.

structure where the diffusion of oxygen ions and the consequent formation of oxygen vacancies happen globally across the interface, the formation of oxygen vacancies at the interface is limited in the Ta/G/Ta_2O_5 device. In this case, both the vertical diffusion through the graphene layer and the lateral diffusion afterwards inside the Ta_2O_5 switching layer can contribute to the slow change in the device conductance, schematically shown in figure 5.7e. It should be noted that annealing devices with a reactive layer can lead to forming-free RS behaviors since the ability to generate

V_Os in the switching layer during annealing eliminates the need for the forming process. [19] However, most Ta/Ta_2O_5 devices could not be erased after annealing. As shown in the inset of Figure 5.7b, even if a RESET voltage is applied first without any SET process, an increase in device current is still observed leading to a complete short. Only a few Ta/Ta_2O_5 devices after annealing showed repeatable switching behavior. Even in these devices a very conductive HRS is observed compared to the unannealed devices, shown in Figure 5.7c. These effects can be explained by the global injection of V_Os during annealing (Figure 5.7d), which causes the Ta_2O_5 switching layer to be flooded with V_Os and results in leaky HRS and SA1 failures. On the other hand, for the Ta/G/Ta_2O_5 devices, no significant changes in the HRS and LRS states were observed after annealing, and reliable forming-free behaviors can be obtained in the devices after annealing, as shown in Figures 5.7b and c. We believe continued optimization of the local seeding effect through nanopores in graphene can provide a systematic design methodology to control and modulate ionic transport in memristive devices and achieve improved device performance including programming current, device variability and reliable forming-free behaviors, which are highly desirable for large scale implementations in RRAM and neuromorphic computing applications.

5.8 Conclusion

In summary, we show that ionic transport in oxide-based memristive devices can be controlled by a graphene ion-blocking layer with engineered nanopores. The migration of oxygen ions and associated redox reactions are blocked by the inserted graphene layer, and oxygen vacancy injection into the switching layer is restricted and only occurs through localized openings in the graphene layer. This local seeding effect through the graphene layer leads to improved

device performance and allows systematic tuning of the RS effects. The effects of graphene at the interface between a reactive metal and the switching layer were verified by high-resolution microscopy and spectroscopy, and supported by first-principles simulations. By controlling the internal ionic dynamic at the atomic scale, this approach provides an attractive option to overcome existing challenges in memristive device research and can possibly lead to large-size systems for different applications.

5.9 References

1. Kim, K. M., Jeong, D. S. & Hwang, C. S. Nanofilamentary resistive switching in binary oxide system; a review on the present status and outlook. *Nanotechnology* **22,** 254002–18 (2011).

2. Waser, R., Dittmann, R., Staikov, G. & Szot, K. Redox-Based Resistive Switching Memories – Nanoionic Mechanisms, Prospects, and Challenges. *Advanced Materials* **21,** 2632–2663 (2009).

3. Lee, M.-J. *et al.* A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O5−x/TaO2−x bilayer structures. *Nature Materials* **10,** 625–630 (2011).

4. Lee, H. Y. *et al.* Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO2 based RRAM. in 1–4 (IEEE, 2008). doi:10.1109/iedm.2008.4796677

5. Lee, J., Du, C., Sun, K., Kioupakis, E. & Lu, W. D. Tuning Ionic Transport in Memristive Devices by Graphene with Engineered Nanopores. *ACS Nano* **10,** 3571–3579 (2016).

6. Bunch, J. S. *et al.* Impermeable Atomic Membranes from Graphene Sheets. *Nano Letters* **8,** 2458–2462 (2008).

7. Garaj, S. *et al.* Graphene as a subnanometre trans-electrode membrane. *Nature* **467,** 190–193 (2010).

8. Lee, J. H. *et al.* Wafer-Scale Growth of Single-Crystal Monolayer Graphene on Reusable Hydrogen-Terminated Germanium. *Science* **344,** 286–289 (2014).

9. Tan, C. *et al.* Recent Advances in Ultrathin Two-Dimensional Nanomaterials. *Chem. Rev.* **117,** 6225–6331 (2017).

10. Schneider, G. F. *et al.* DNA Translocation through Graphene Nanopores. *Nano Letters* **10,** 3163–3167 (2010).

11. Lübben, M. *et al.* Graphene-Modified Interface Controls Transition from VCM to ECM Switching Modes in Ta/TaO xBased Memristive Devices. *Advanced Materials* **27,** 6202–6207 (2015).

12. Wedig, A. *et al.* Nanoscale cation motion in TaOx, HfOx and TiOx memristive systems. *Nature Nanotechnology* **11,** 67–74 (2015).

13. Goux, L. *et al.* Role of the Ta scavenger electrode in the excellent switching control and reliability of a scalable low-current operated TiN\Ta<inf>2</inf>O<inf>5</inf>\Ta RRAM device. in 1–2 (IEEE, 2014). doi:10.1109/vlsit.2014.6894401

14. Kim, H., McIntyre, P. C., On Chui, C., Saraswat, K. C. & Stemmer, S. Engineering chemically abrupt high-k metal oxide⁄silicon interfaces using an oxygen-gettering metal overlayer. *Journal of Applied Physics* **96,** 3467–3472 (2004).

15. Fleming, R. M. *et al.* Defect dominated charge transport in amorphous Ta2O5 thin films. *Journal of Applied Physics* **88,** 850–862 (2000).

16. Kim, S., Choi, S. & Lu, W. Comprehensive physical model of dynamic resistive switching in an oxide memristor. *ACS Nano* **8,** 2369–2376 (2014).

17. O'Hern, S. C. *et al.* Selective Molecular Transport through Intrinsic Defects in a Single Layer of CVD Graphene. *ACS Nano* **6,** 10130–10138 (2012).

18. Abbas, A. N. *et al.* Patterning, Characterization, and Chemical Sensing Applications of Graphene Nanoribbon Arrays Down to 5 nm Using Helium Ion Beam Lithography. *ACS Nano* **8,** 1538–1546 (2014).

19. Lohn, A. J., Stevens, J. E., Mickel, P. R., Hughart, D. & Marinella, M. J. A CMOS Compatible, Forming Free TaOx ReRAM. *ECS Transactions* **58,** 59–65 (2013).

Chapter 6 K-means Clustering Analysis using a Memristor Neural Network

6.1 Introduction

Neuromorphic computing systems based on emerging devices such as memristors have attracted great interest, especially following recent advances of experimental demonstrations at the network-level aimed for practical tasks[1-7]. As a hardware system that offers co-location of memory and logic and highly parallel processing capability, memristor-based networks enable efficient implementation of machine learning algorithms[8-10]. Among the approaches, unsupervised learning is of growing importance since it relies on (abundant) unlabeled training dataset[11], which is far cheaper to obtain than those required by supervised learning algorithms[12-14]. During training, unsupervised learning rules rely on indicators of similarity between the input feature vectors and the learned feature vectors (dictionary elements), e.g. distance between these vectors in Euclid space, to identify the dictionary element that best matches the input and subsequently adjust the weights accordingly[15-17]. If the dictionary elements are normalized, finding the shortest Euclidean distance is equivalent to finding the smallest dot-product between the input vector and the dictionary element vector, which can be readily obtained in a memristor crossbar[1,2,4]. However, when the dictionary elements cannot be normalized during learning, finding the shortest Euclidean distance is no longer trivial, and can cause significant overhead in hardware implementation.

In this study, we propose and experimentally demonstrate an approach that directly compares the shortest Euclidean distance in a memristor crossbar hardware system, without normalizing the weights. As a test case, we use this approach to implement the *K*-means

algorithm[15], one of the most widely used unsupervised methods in cluster analysis[18,19], experimentally in a memristor crossbar array. The network successfully performs cluster analysis through online training for different kinds of inputs regardless of the initial centroid locations.

6.2 Methodology

Device fabrication. The Ta_2O_5-based memristor networks used in this study were fabricated in a crossbar structure in which a memristive device is formed at each crosspoint. Starting from a SiO_2 (100 nm) / Si substrate, the bottom Pd electrodes (50 nm thickness) with an adhesion layer of NiCr (5 nm thickness) were formed by photolithography (GCA AS200 Autostep), e-beam evaporation and lift-off processes. The Ta_2O_{5-x} film (3.5 nm thickness) was then deposited by radio frequency (RF) sputtering with very low power (30 W) using a Ta_2O_5 ceramic target at room temperature. Next, the 40 and 40/100 nm thick Ta and Pd/Au top electrodes were patterned in the same manner as the bottom electrodes and deposited by direct current (DC) sputtering with power of 100 W and e-beam evaporation, respectively. Finally, the bottom electrode pads were opened by etching the Ta_2O_{5-x} layer using a reactive ion etching (RIE) process with SF_6/Ar gases. After fabrication, the memristor chip was wire-bonded to a chip carrier and integrated with the test board.

Array test board set-up. A custom-built test board was used for the experiments. The board can apply arbitrary voltage pulses to the selected rows and columns and measure output currents from any selected row or column. Two digital-to-analog converters (DACs) are used as voltage sources (from -5V to +5V) for the rows, and similar setups are sued for the columns of the memristor array. The signals are applied through a switching matrix to the selected addresses. DAC0 (DAC2) is connected to the selected rows (columns) during programming, and DAC1 (DAC3) is connected

to the remaining unselected rows (columns) to supply the protective voltage during programming. During read operation, the output node is connected to the Operational Amplifier, with the other unselected rows and columns grounded. The board is programmed by an FPGA (Xilinx, Spartan6) that is integrated on the test board. The codes are executed in a fully automated manner through Python programming.

6.3 Mapping *K*-means onto RRAM Crossbar Network

The *K*-means[15] algorithm aims to partition a set of vector inputs into *K* clusters through exploratory data analysis, as schematically shown in Figure 6.1a (for *K*=3). From the random initial centroid locations, the network evolves as input datapoints are assigned to different clusters based on the distances between the input datapoint and the different centroid locations, followed by updating the centroid locations of the clusters. With a relatively simple form, *K*-means provides a comparable solution to more complex approaches such as autoencoders[21] for pre-clustering of unlabeled dataset through online training, and reduces the original input space into disjoint smaller

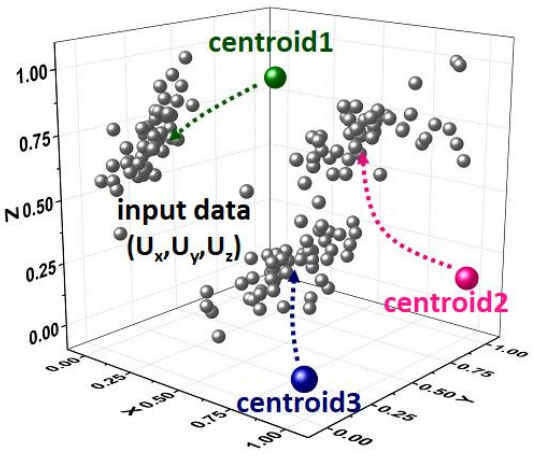

Figure 6.1: Schematic of the *K*-means algorithm showing the evolution of the *K* centroid locations during online learning.

sub-spaces for subsequent use of fine clustering algorithms or data classification through another supervised layer[22,23].

However, there are two challenges for experimentally implementing algorithms such as *K*-means in memristor-based systems. The first challenge is obtaining the Euclidean distance directly in hardware. Memristor arrays can readily implement vector-vector dot-product operations[1,2,24,25]. However, when the weights are not normalized, the feature vector that produces the largest dot-product with the input vector is not necessarily the one having the shortest distance to the input, as will be discussed in detail below. The second challenge is related to the fact that *K*-means is not aimed at minimizing an error/cost. In previously demonstrated systems, the operation of the memristor network is to minimize a cost function based on the learning rule - either the output label error for supervised learning[1,3] or the cost function for representing the input in certain unsupervised learning cases[2]. As a result, the algorithm relies on a known reference (either the output label or the original input) to calculate the error/cost. The reference in turn provides a feedback mechanism that helps the network converge near the targeted solution, even in the presence of sizeable device variations[26,27]. This feedback mechanism however is missing in many unsupervised learning and arithmetic operation processes, including the *K*-means approach, since such algorithms do not explicitly aim to minimize an error against a reference, and thus hardware demonstration will post more stringent requirement on the device performance.

6.4 Extended RRAM Network

Here we address these challenges and experimentally implement the *K*-means algorithm in memristor crossbar array-based hardware as a test case. In the implementation, locations of the *K* centroids are directly mapped as weights in the memristor array in column-wise fashion, i.e., W_{mn}

at cross-point (m, n) in the memristor array corresponds to the m-th coordinate value of the n^{th} centroid. The weights are in turn represented by the state variable (w) in the device model, which is linearly mapped onto the device conductance. With the input vector set U and a randomly generated initial weight matrix W, the K-means algorithm iteratively performs two successive operations to assign the inputs to the appropriate clusters. The first process, the searching step, is to find the nearest centroid for a given input and assign the input to the associated cluster. It is then followed by the second process, the update (learning) step, which updates the selected centroid coordinates due to changes in the cluster composition. The crossbar structure as shown in Figure 6.2 is extremely efficient in implementing vector-matrix multiplication operation, which is one of the core operations[1-4] in neuromorphic systems, through simple Ohm's raw and Kirchhoff's law, i.e., the current (or charge) collected at each output neuron ($I_n = \sum_j V_j \cdot W_{jn}$) represents the dot-product of the input voltage vector and the stored conductance vector. In general vector-vector dot-product operation provides a good indication of the similarity between the input vector and the stored vector, and thus can be viewed as performing a pattern-matching operation and is commonly used in machine-learning algorithms, particularly if the stored feature vectors can be readily normalized. However, algorithms such as K-means rely on finding the exact Euclidean distances, not just the similarity between the vectors, to decide the winning neurons and perform the updates.

The Euclidean distance of the input vector U and the weight vector W_n for the n^{th} centroid is determined by equation (6.1):

$$\|U - W_n\|^2 = U^2 - 2U \cdot W_n + {W_n}^2 \tag{6.1}$$

Beside the dot-product term $U \cdot W_n$, two other terms, U^2 and W_n^2, are required to obtain the Euclidean distance $\|U - W_n\|$. The U^2 term depends on the input only and thus is a constant for all centroids and will not affect the comparison when determining the winning neuron. However,

the $W_n^2 = \sum_j W_{jn}^2$ term, representing the L2 norm of the weight vector associated with centroid *n*, can in general be different for each centroid. Only in certain conditions will the weight vector be naturally normalized (e.g. when the weight update follows the Oja's rule[27]). Numerically normalizing the weight vectors is expensive, and if not performed, the dot-product $U \cdot W_n$ will not in general correctly represent the Euclidean distance due to differences in the W_n^2 term.

Below we show that Euclidean distance comparisons can still be obtained in memristor crossbar-based implementations. We note that if the array matrix is expanded to include one additional row representing the W^2 term (called the W^2 scheme approach for convenience), results from equation (6.1) can still be obtained in the memristor array during a single read operation using the same vector-matrix multiplication approach. Here, the new matrix consists of both the original *W* matrix (sized $M \times N$) and a new *S* matrix (sized $1 \times N$) (Figure 6.2). The *S* matrix in this case is just a single row and stores the average value of the squared weight ($<W_n^2> = \sum_j (\frac{W_{jn}^2}{M})$) for the n^{th} centroid. The key is then to obtain the desired output using this expanded matrix and to allow the *S* matrix to be updated correctly during unsupervised online learning.

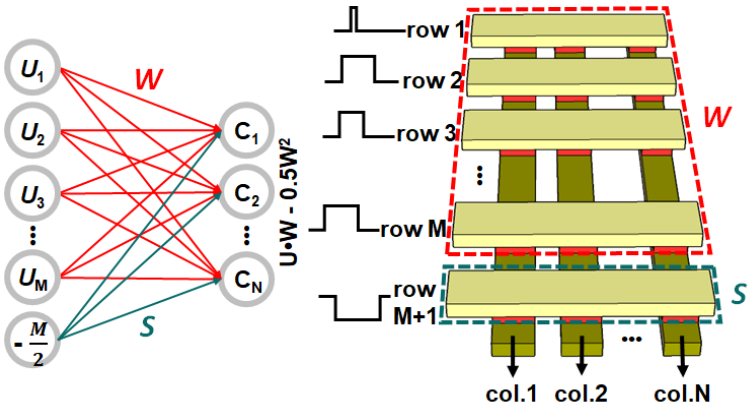

Figure 6.2: Mapping the proposed algorithm onto the memristor array. The coordinates of a centroid is stored as the memristor conductance values in the corresponding column in the *W*-matrix. The W^2 information for the centroid is stored by the memristor conductance in the S matrix in the same column. The input values are coded as pulses with different widths and are applied to the rows of the expanded *W*-matrix. The accumulated charges at the columns' outputs allow direct comparison of the Euclidean distances.

During Euclidean distance calculation, the input vector will now include the original input *U* (applied to the rows of the original *W* matrix), and an additional, constant element -*M*/2 that will be applied to the *S* matrix, as shown in Figure 6.2. Here *M* is the number of rows in the *W* matrix. The choice of the input and the *S* matrix is to ensure the values of the elements in the *S* matrix are in the same range as the values in the *W* matrix, so that the same type of memristor devices can be used for the expanded matrix. Experimentally, the elements in the input vector are implemented with voltage pulses having a fixed amplitude (V_{read}=0.3V in our study) and variable widths proportional to the desired input values. The input voltage pulses are applied to the network to perform the vector-matrix multiplication, and the charge Q_n accumulated at each output will represent the distance between the original input vector *U* and the weight vector W_n since $Q_n = \sum_j (U_j \cdot W_{jn} - \frac{W_{jn}^2}{2})$ differs from the exact expression (1) only by a (-) sign and a constant U^2. As a result, the higher the output charge obtained from a particular column in the memristor array, the shorter the distance is between the input *U* and the centroid represented by the column.

After identifying the nearest centroid, the second step is to update both the weights associated with the centroid (representing the coordinates of the centroid location) in the original *W* matrix and the *S* matrix representing the $<W^2>$ term. We have developed a learning rule that can be readily implemented in the memristor hardware, shown in equation (6.2).

$$\Delta W_n = \eta \cdot (input - W_n) \tag{6.2a}$$

$$\Delta S_n = \sum_j (\frac{W_{jn}^2}{M}) - S_n \tag{6.2b}$$

Here η is a constant and represents the learning rate for the *W* matrix. Note update is only performed for the column corresponding to the nearest centroid, and both the *W* matrix and the *S* matrix are updated simultaneously. By repeating the iterative training process with the searching step and the updating step, the network stabilizes and learns the centroids of the *K* clusters.

6.5 Optimization of Analog Switching Behaviors

As discussed in Section 6.3, experimental implementation of the *K*-means algorithm in RRAM networks requires more accurate weight update based on better analog switching behaviors. However, conventional Ta_2O_5-based devices show highly nonlinear analog switching effects, as shown in Figure 6.3b. Specifically, when the device is stimulated with a series of identical programming pulses, the device switching characteristics can be roughly divided into two regimes. The first regime exhibits abrupt conductance changes with a relatively larger dynamic range, while the second regime shows more gradual conductance change, but with a very small dynamic range. From device simulation (Figure 6.3c), we found that the initial abrupt change is caused by the vertical growth of the filament (marked as states A,B, and C in Figure 6.3b and c)

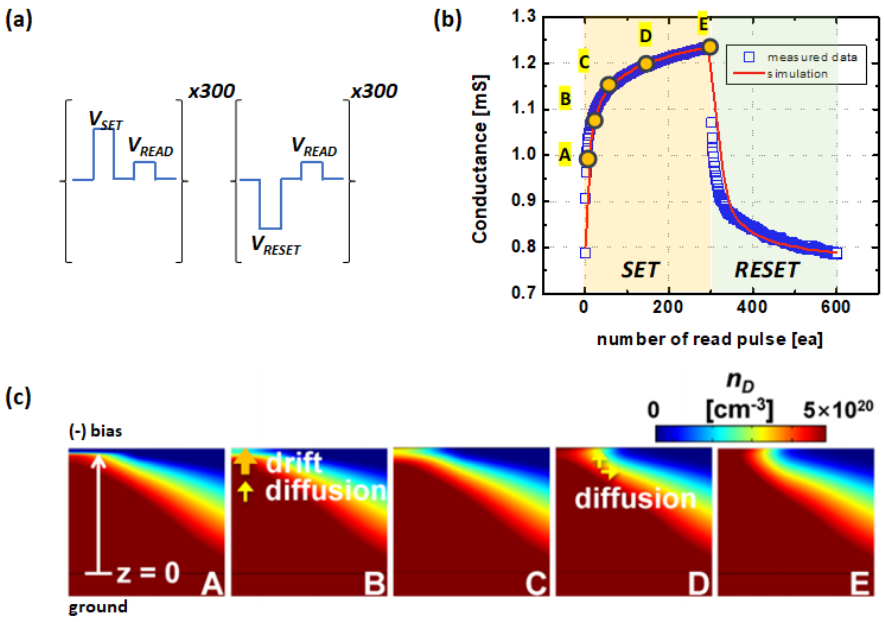

Figure 6.3: (a) Pulse trains for analog switching (b) Analog conductance update characteristics, showing initial abrupt change followed by gradual saturation. (c) Simulation results showing oxygen vacancy profile during the set process.[15]

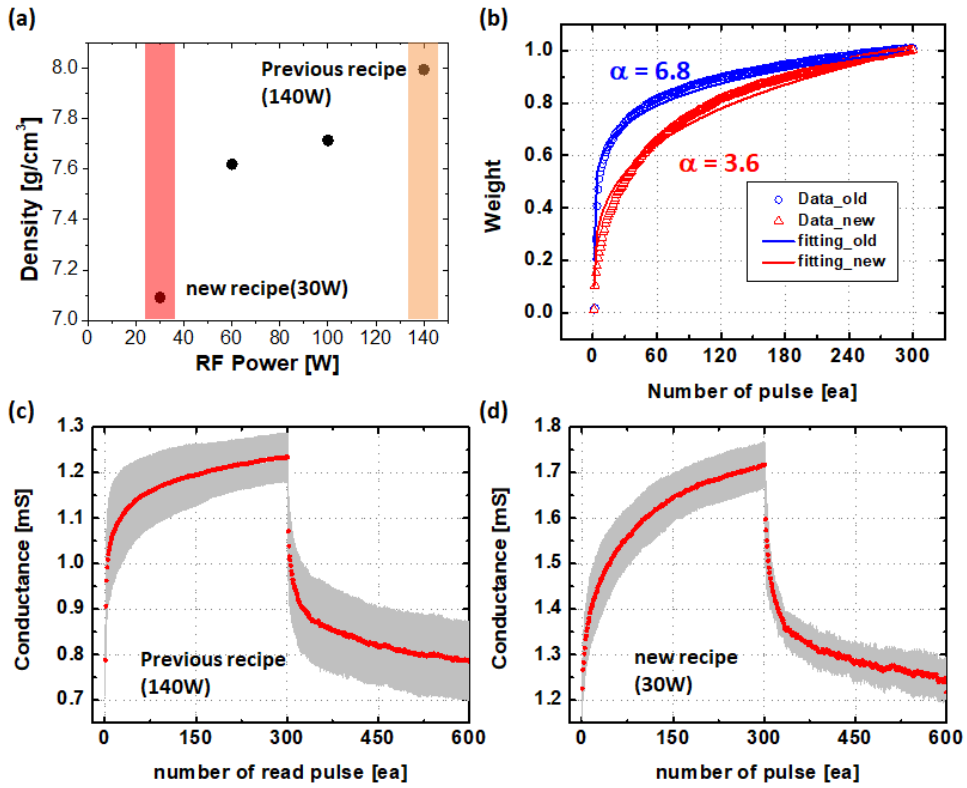

Figure 6.4: (a) Film density of Ta_2O_5 films deposited with different RF sputtering powers (b) Comparison of the analog conductance update characteristics of the high- and low-density devices, showing improved linearity in the low-density device (c-d) Analog conductance update characteristics of a device made by high-density (c) and low-density (d) film.

and the second regime with more gradual conductance change corresponds to the lateral growth of the filament through V_O diffusion (marked as states D and E).

To improve the gradual conductance change characteristics and the dynamic range, we employed a low-density Ta_2O_5 film as the switching layer based on the assumption that a film with lower density offers lower V_O diffusion barrier and will lead to more lateral growth and larger dynamic range. We note that a low-density film could be deposited by using low RF power during the sputtering process, as shown in Figure 6.4a. The smaller voltage drop across plasma sheath in the low-power condition reduces the kinetic energy of energetic ions and thus the energy of the sputtered materials. As a result, packing factor will be reduced, leading to a low-density amorphous

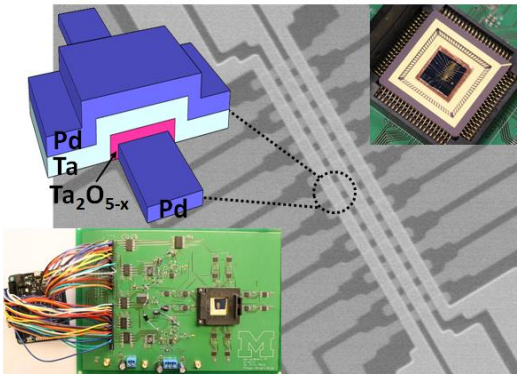

Figure 6.5: Scanning electron micrograph (SEM) image of a fabricated crossbar array used in this study. Upper left inset: schematic of the device structure. Lower left inset: Photo of the test board. The memristor array is wire-bonded (upper right inset) and integrated into the board system.

film. Indeed, a device with the low-density film exhibits a more incremental switching with improved dynamic range, as shown in Figure 6.4c and d. The linearity of the analog switching can be determined by fitting the analog conductance update curve with equation:

$$w = (A \times \Delta t + B)^{\frac{1}{\alpha}} \tag{6.3}$$

where A, B, and Δt are fitting parameters and the applied pulse width (or the number of applied pulses). The optimized device with the low-density film exhibits a reduced linearity factor, α of 3.6, compared to the control device with α of 6.8. The lower α suggests a more linear conductance update process (with α = 1 for the ideal linear response case) (Figure 4.4b). Based on the optimized Ta_2O_{5-x} RRAM device, experimental implementation of the K-mean algorithm was performed using the as-fabricated 16×3 crossbar array, as shown in Figure 6.5.

6.6 Experimental Implementation of the *K*-means Algorithm

The *K*-means algorithm was experimentally implemented using the memristor array and a custom-built testing board. The test board allows arbitrary pulse signals to be sent to and electronic

current collected from either individual devices or multiple devices in multiple rows and columns simultaneously in parallel.

A simple 2-dimensional (2D) dataset was first used to test the system. Specifically, 50 2D data points that can be explicitly partitioned into three clusters, i.e. K=3, were manually generated in the study to verify the operation of the memristor network. Figure 6.6a shows the evolution of the learned centroid positions obtained from the memristor-based K-means system, using online unsupervised training for three different initial weights conditions. During training, the closest

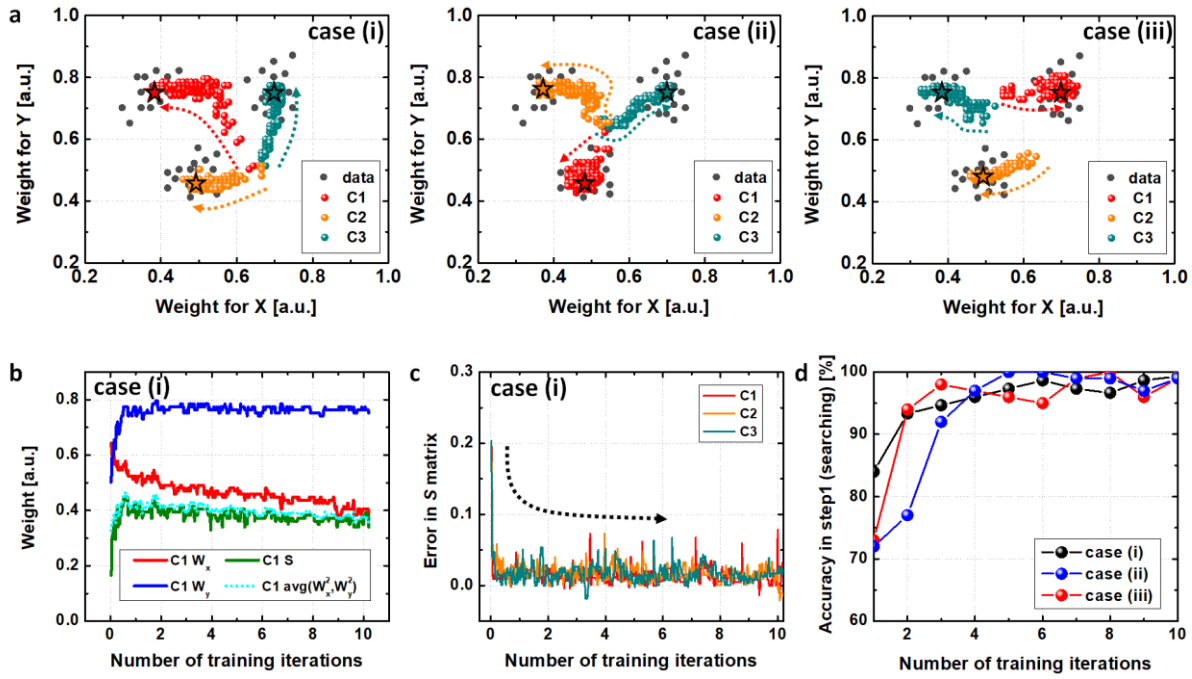

Figure 6.6: Experimental implementation of K-means using the memristor crossbar. a, Evolution of the three centroid locations during training, for three cases with different initial configurations: i) all three centroids were located at the same position outside the dataset; ii) all three centroids were located at the same position inside the dataset; iii) randomly assigned initial positions. The final locations of the centroids are represented by the stars. b, Evolution of the W elements W_x, and W_y (representing the centroid coordinates) for centroid 1 in case I, along with the centroid's S element, and a reference showing the calculated $<W^2>$ during training. Each iteration includes 50 training operations for a given input data set. c, Difference between the stored S element and the calculated $<W^2>$ for the three centroids, showing a rapid decrease after only a few training steps. d, Success rate of correctly finding the nearest centroid using the memristor network, as a function of the iteration number during training. The success rate becomes > 95% after 4 iterations.

centroid to an input is first determined from the memristor array using the W^2 approach, followed by updates of the *W* and *S* elements for the selected centroid based on equation (6.2). For example in case (i), the three centroids were initially placed at the same location, *i.e.* having nearly identical *W* vectors in the memristor matrix (with the small differences due to device variations among the different columns). During training centroid 2 moved in the left direction towards one group, while centroid 3 moved in the up direction. Centroid 1 moved in the upper left direction with relatively large movements at the beginning of training, and turned left towards the group of data farthest from the initial point. After training, the three centroids settled to the centers of the respective clusters, and every point in the dataset can be properly assigned to one of the three clusters. Other initial centroid locations, such as having all centroids located in the center of the dataset (case ii), and having the centroids randomly distributed (case iii) have also been tested, and the memristor network can successfully perform *K*-means clustering in all cases (Figure 6.6a) using our experimental setup, demonstrating the reliability of the proposed training algorithm and the hardware implementation using physical memristor crossbar arrays, even with non-normalized weights.

We note that the reliability of the *K*-means algorithm implementation depends critically on how accurately the *W* and *S* elements are modulated during training, since errors in these elements directly affect the obtained Euclidean distance and the subsequent identification of the winning centroid. The experimentally obtained evolution of the *W* and *S* elements during training are shown in Figure 6.6b. One can see that as a *W* vector (e.g., W_x and W_y of centroid 1 in Case (i) shown in Figure 6.6b) is updated when a new data point is added to the cluster, the corresponding *S* element is properly adjusted based on the proposed algorithm, and the learned *S* elements in the memristor network following the proposed equation (6.2b) are indeed in good agreement with the calculated

value of $<W^2>$. Specifically, Figure 6.6c shows the difference between the stored *S* values and the calculated $<W^2>$ value (numerically calculated from the measured values of the *W* matrix). The error is large in the beginning because the *S* elements were not initialized according to the *W* matrix values. Importantly, the error drops rapidly with only a few training steps and remains very low, due to the excellent incremental conductance modulation capability of the physical device as shown in Figure 6.4c. As a result, the system can correctly find the nearest centroid with over 95% success rate after 4 iterations as shown in Figure 6.6d, verifying that the proposed W^2 scheme can be used to efficiently calculate distances between vectors without compute-intensive normalization processes.

6.7 Conclusion

In this work, we experimentally demonstrated memristor-based neural networks for *K*-means clustering analysis. A W^2 scheme was proposed to allow accurate calculation of the Euclidean distance, which is an essential operation in many machine learning algorithms, through direct vector-matrix multiplication in memristor networks with non-normalized weights. Without a cost function that provides a feedback mechanism, the *K*-means algorithm poses stricter requirements on compute accuracy and device variability compared with other neuromorphic computing algorithms. With an expanded weight matrix, properly designed training rules, and improved device properties, we show *K*-means clustering can be reliably implemented using memristor networks. With continued device and algorithm optimizations, the results obtained here can pave the way towards practical memristor network hardware for more broad applications beyond neuromorphic systems.

6.8 References

1. Prezioso, M. *et al.* Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521,** 61–64 (2015).

2. Sheridan, P. M. *et al.* Sparse coding with memristor networks. *Nature Nanotechnology* **12,** 784–789 (2017).

3. Yao, P. *et al.* Face classification using electronic synapses. *Nature Communications* **8,** 15199 (2017).

4. Choi, S., Shin, J. H., Lee, J., Sheridan, P. & Lu, W. D. Experimental Demonstration of Feature Extraction and Dimensionality Reduction Using Memristor Networks. *Nano Letters* **17,** 3113–3118 (2017).

5. Strukov, D. B., Snider, G. S., Stewart, D. R. & Williams, R. S. The missing memristor found. *Nature* **453,** 80–83 (2008).

6. Waser, R. & Aono, M. Nanoionics-based resistive switching memories. *Nature Materials* **6,** 833–840 (2007).

7. Jo, S. H. *et al.* Nanoscale Memristor Device as Synapse in Neuromorphic Systems. *Nano Letters* **10,** 1297–1301 (2010).

8. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529,** 484–489 (2016).

9. Merolla, P. A. *et al.* A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345,** 668–673 (2014).

10. Indiveri, G. & Liu, S.-C. Memory and Information Processing in Neuromorphic Systems. *Proc. IEEE* **103,** 1379–1397 (2015).

11. Hinton, G. E. & Sejnowski, T. J. Unsupervised learning: foundations of neural computation. Computers & Mathematics with Applications **38,** (1999).

12. Goodfellow, I. J. *et al.* Generative adversarial nets. in **3,** 2672–2680 (2014).

13. Hinton, G. E., Osindero, S. & Teh, Y.-W. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation* **18,** 1527–1554 (2006).

14. Hinton, G. E. Reducing the Dimensionality of Data with Neural Networks. *Science* **313,** 504–507 (2006).

15. MacQueen, J. Some methods for classification and analysis of multivariate observations. in (1967).

16. Kohonen, T. The self-organizing map. *Neurocomputing* **21,** 1–6 (1998).

17. Filippone, M., Camastra, F., Masulli, F. & Rovetta, S. A survey of kernel and spectral methods for clustering. *Pattern Recognition* **41,** 176–190 (2008).

18. Jiang, Y., Kang, J. & Wang, X. RRAM-based parallel computing architecture using k-nearest neighbor classification for pattern recognition. *Scientific Reports* **7,** 45233 (2017).

19. Jain, A. K., Murty, M. N. & Flynn, P. J. Data clustering: a review. *ACM Comput. Surv.* **31,** 264–323 (1999).

20. Bache, K. & Lichman, M. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2013. (URL: http://archive. ics. uci. edu/ml, 2013).

21. Bengio, Y. Learning Deep Architectures for AI. *FNT in Machine Learning* **2,** 1–127 (2009).

22. Lin, D. & Wu, X. Phrase clustering for discriminative learning. in **2,** 1030 (Association for Computational Linguistics, 2009).

23. Coates, A. & Ng, A. Y. in *Neural Networks: Tricks of the Trade* **7700,** 561–580 (Springer, Berlin, Heidelberg, 2012).

24. Yang, J. J., Strukov, D. B. & Stewart, D. R. Memristive devices for computing. *Nature Nanotechnology* **8,** 13–24 (2013).

25. Merced-Grafals, E. J., Davila, N., Ge, N., Williams, R. S. & Strachan, J. P. Repeatable, accurate, and high speed multi-level programming of memristor 1T1R arrays for power efficient analog computing applications. *Nanotechnology* **27,** 365202 (2016).

26. Yu, S. *et al.* A Low Energy Oxide-Based Electronic Synaptic Device for Neuromorphic Visual Systems with Tolerance to Device Variation. *Advanced Materials* **25,** 1774–1779 (2013).

27. Sheridan, P. M., Du, C. & Lu, W. D. Feature Extraction Using Memristor Networks. *IEEE Transactions on Neural Networks and Learning Systems* **27,** 2327–2336 (2016).

Chapter 7 Future Works

7.1 Large RRAM Crossbar Array

In Chapter 6, we demonstrated the experimental implementation of data clustering algorithm using a RRAM crossbar array. However, larger crossbar arrays are still needed to implement more complex or deeper networks. When the size of a crossbar array is increased, several issues are expected. First, line resistance of bottom and top electrodes will increase, leading to severe voltage dividing effects and thus large variations in the actual applied field at each device depending on the location of devices in a crossbar array. To reduce line resistance, optimization of the electrode materials and structures, such as thickness of bottom electrode, a side-wall spacer for the mitigation of the ridges formed around the bottom electrode edges, and buried electrode need to be systematically investigated. Additionally, it is highly desirable to reduce the operation current level of oxide-based RRAM devices because it will mitigate the effects of line resistance and also reduce power consumption.

Another important task to achieve fully functional, large passive crossbar array is incorporating a component that can work as a current compliance. Even for devices having electroforming-free behavior (or with low enough forming voltage), the current during the first several programming cycles still needs to be limited to prevent damage to the device. In a small-size array, external resistance (e.g., from extension lines) can limit current flow by adding an appropriate amplitude of series resistance as the array used for the *K*-means clustering implementation in Chapter 6. However, the number of parallelly connected RRAM devices increases quadratically as array size increases, leading to a very severe voltage divider effect at the

external resistance. As a result, higher voltage will be required to subsequently form/program devices in the array, damaging devices that have already been formed earlier.

One solution is to add on-cell resistance instead of external resistance. An additional layer added in a RRAM device stack can provide a constant series resistance as an on-cell resistor regardless of how many devices are already formed/programmed in an array. For example, ITO thin films with different oxygen flow during sputtering deposition can result in controlled resistivity values in a large range of over four order of magnitude. By finding proper oxygen flow and thickness of an ITO layer, RRAM devices with self-compliance layers could be achieved.

Lastly, device variability and yield still need to be improved further to achieve large crossbar arrays. Specifically, RRAM devices are inherently stochastic during the resistive-switching process, originating from the stochastic, thermally activated ion migration and redox processes. These stochastic effects lead to temporal variations, in which the switching time or switching voltage is broadly distributed from cycle-to-cycle, even for the same device. The stochastic growth of the filament, combined with the disordered nature of the amorphous switching material, also results in large device-to-device variations. Confining a filament would be a good approach to solve these issues. The concentrated electric field resulting from engineered nanostructures and optimized switching materials can help guide ion transport and redox reactions, thus suppressing random and uncontrolled filament growth. Further research will be needed to develop practical methods and structure for large arrays.